

ITI0209: User Interfaces

13. Introduction to Data Visualization

Martin Verrev

Spring 2024

John W. Tukey

EXPLORATORY DATA ANALYSIS



"The greatest value of a picture is when it forces us to notice what we never expected to see."

John Tukey, Exploratory Data Analysis

Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812-1813.

Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite Paris, le 20 Novembre 1869.

Les nombres d'hommes présents sont représentés par les largeurs des zones colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zones. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M. M. Chiers, de Légar, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.

Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.

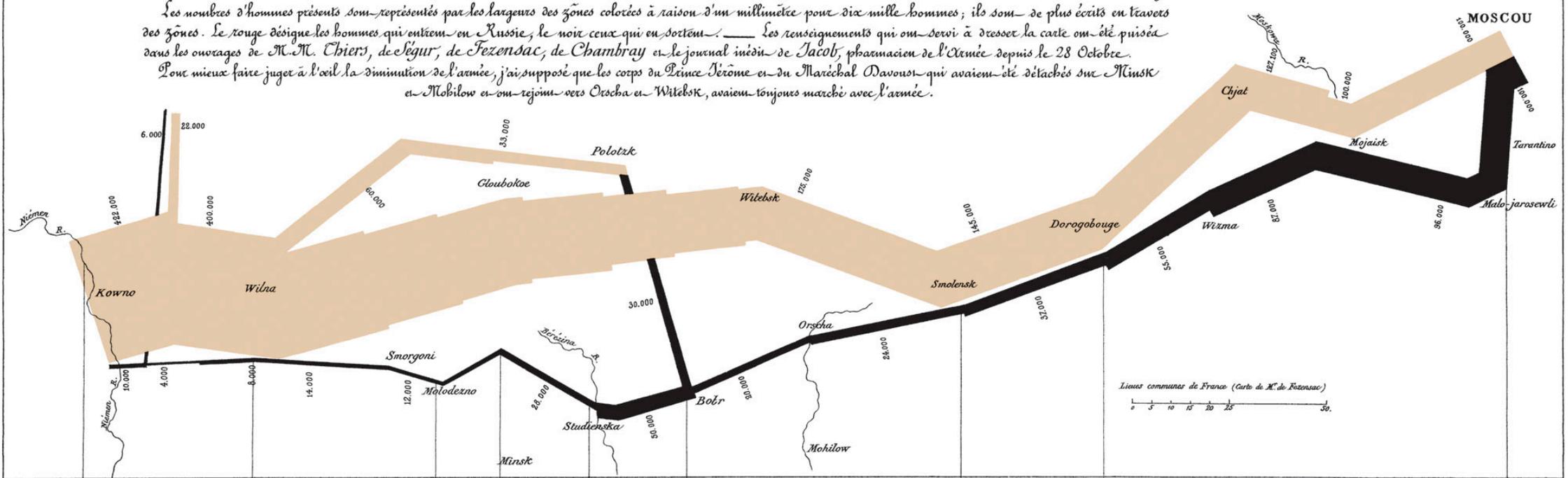
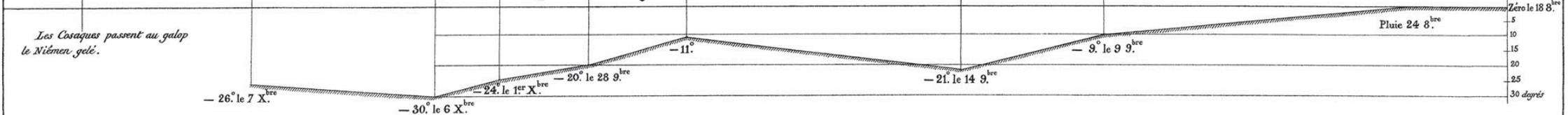


TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.

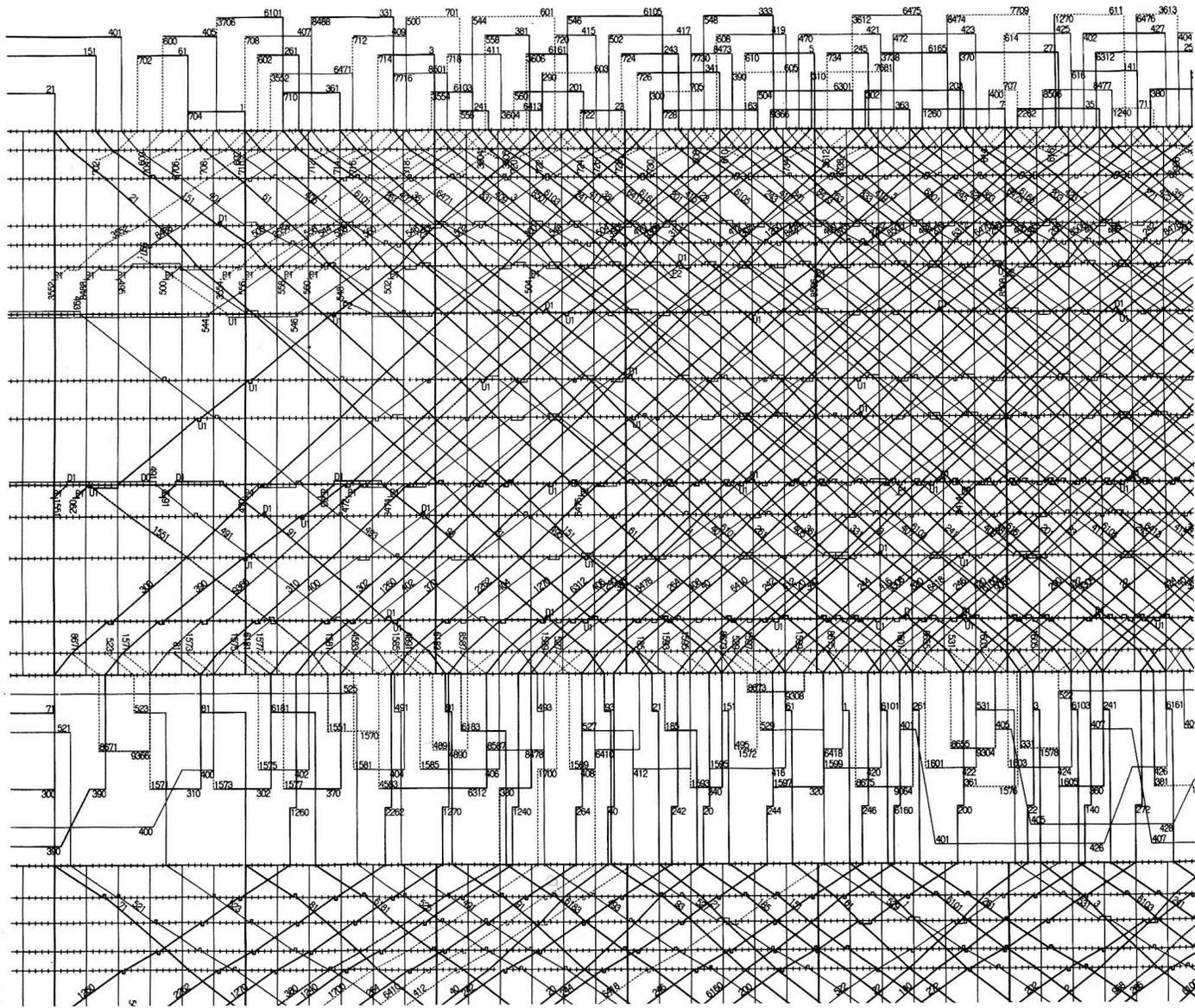


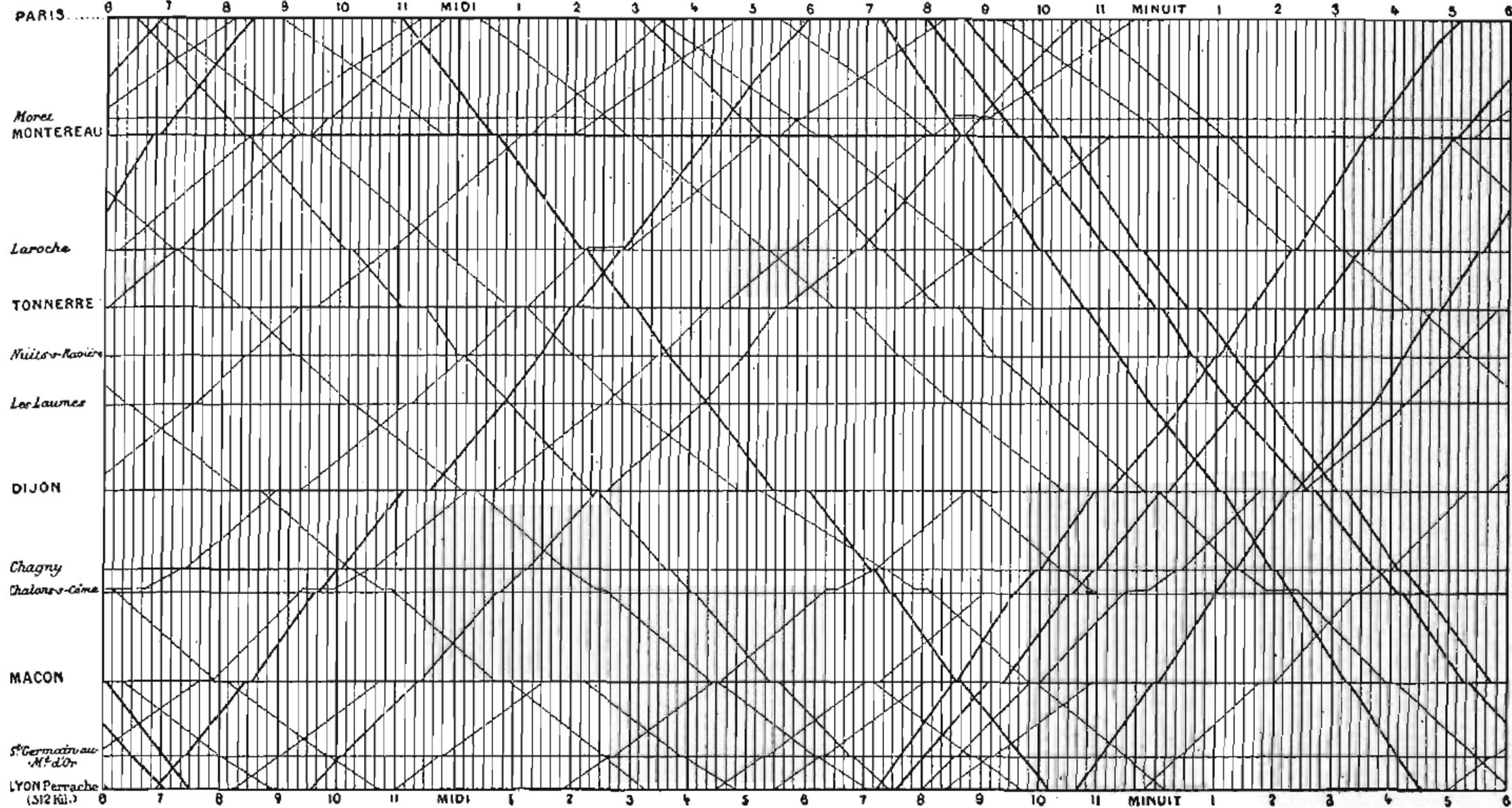
Les Cosaques passent au galop le Niémen gelé.

Autog. par Regnier, 8. Par. S^{te} Marie S^t O^u à Paris.

Imp. Lith. Regnier et Douard.

<https://www.esri.com/content/dam/esrisites/en-us/esri-press/pdfs/mapping-time-illustrated-minards-map-napoleons-russian-campaign-1812-sample-chapter.pdf>





E. J. Marey, *La méthode graphique* (Paris, 1885), p. 20. The method is attributed to the French engineer, Ibry.

Visualizing information is important

The brain doesn't just process information that comes through the eyes. It also creates mental visual images that allow us to reason and plan actions that facilitate survival. Simply put - a graphical representation is meant to simplify information and make it effortlessly scannable.

The Purpose and Goal

The first and main goal of any graphic and visualization is to be a tool for your eyes and brain to perceive what lies beyond their natural reach.

Check:

<https://informationisbeautiful.net/>

<https://informationisbeautiful.net/beautifulnews/>

<https://www.reddit.com/r/dataisbeautiful>

Why Does it matter?

- **Digestibility:** It makes data easy to digest, turning the daunting into the delightful.
- **Pattern Discovery:** Uncover hidden patterns and correlations within your data, making the complex seem surprisingly simple.
- **Information Compilation:** Imagine having all your information neatly compiled in one place — a visual feast for your analytical appetite.
- **Enhanced Understanding and Retention:** Data visualization not only aids understanding but also boosts retention. Think of it like a well-crafted YouTube video that leaves a lasting impression.

What is Graphical Excellence

1. Well designed presentation of interesting data - a matter of substance, statistics and design
2. Consists of complex ideas communicated with clarity, precision and efficiency
3. That what gives the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space
4. Nearly always multivariate
5. Requires telling the truth about data

Choosing a visualization

There is an infinite number of different visualization types (check <https://d3js.org/>). Handful of types will work for the majority of your needs:

Simple Text. Table. Point. Line. Slopegraph. Bars. Area.

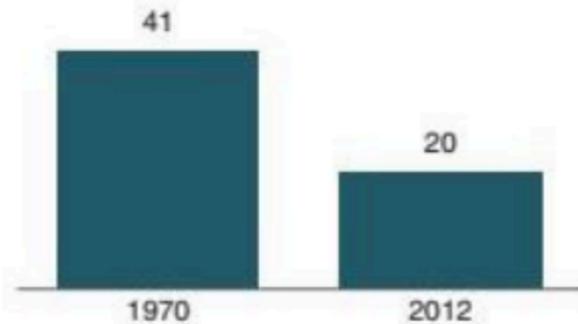


Simple Text

If just a number or two to share

Children with a "Traditional" Stay-at-Home Mother

% of children with a married stay-at-home mother with a working husband



20%

of children had a
traditional stay-at-home mom
in 2012, compared to 41% in 1970

Tables

Tables interact with our verbal system, which means that we read them. If you need to communicate multiple different units of measure it is usually easier with a table than a graph.

Heavy borders

| Group | Metric A | Metric B | Metric C |
|---------|----------|----------|----------|
| Group 1 | \$X.X | Y% | Z,ZZZ |
| Group 2 | \$X.X | Y% | Z,ZZZ |
| Group 3 | \$X.X | Y% | Z,ZZZ |
| Group 4 | \$X.X | Y% | Z,ZZZ |
| Group 5 | \$X.X | Y% | Z,ZZZ |

Light borders

| Group | Metric A | Metric B | Metric C |
|---------|----------|----------|----------|
| Group 1 | \$X.X | Y% | Z,ZZZ |
| Group 2 | \$X.X | Y% | Z,ZZZ |
| Group 3 | \$X.X | Y% | Z,ZZZ |
| Group 4 | \$X.X | Y% | Z,ZZZ |
| Group 5 | \$X.X | Y% | Z,ZZZ |

Minimal borders

| Group | Metric A | Metric B | Metric C |
|---------|----------|----------|----------|
| Group 1 | \$X.X | Y% | Z,ZZZ |
| Group 2 | \$X.X | Y% | Z,ZZZ |
| Group 3 | \$X.X | Y% | Z,ZZZ |
| Group 4 | \$X.X | Y% | Z,ZZZ |
| Group 5 | \$X.X | Y% | Z,ZZZ |

Tables. Heatmap

A heatmap is a way to visualize data in tabular format, where in place of (or in addition to) the numbers, you leverage colored cells that convey the relative magnitude of the numbers.

Table

| | A | B | C |
|------------|-----|-----|-----|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

Heatmap

LOW-HIGH

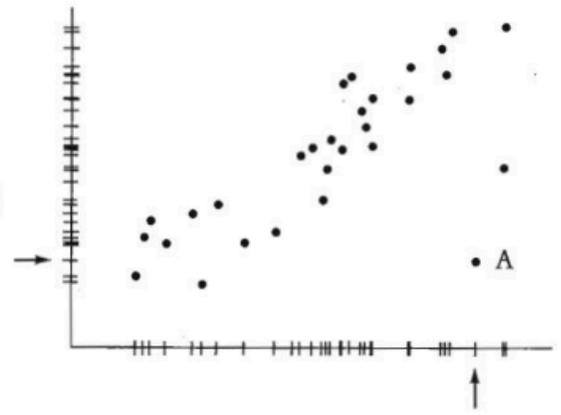
| | A | B | C |
|------------|-----|-----|-----|
| Category 1 | 15% | 22% | 42% |
| Category 2 | 40% | 36% | 20% |
| Category 3 | 35% | 17% | 34% |
| Category 4 | 30% | 29% | 26% |
| Category 5 | 55% | 30% | 58% |
| Category 6 | 11% | 25% | 49% |

Graphs.

While tables interact with our verbal system, graphs interact with our visual system, which is faster at processing information.

| 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
 $r^2 = .67$

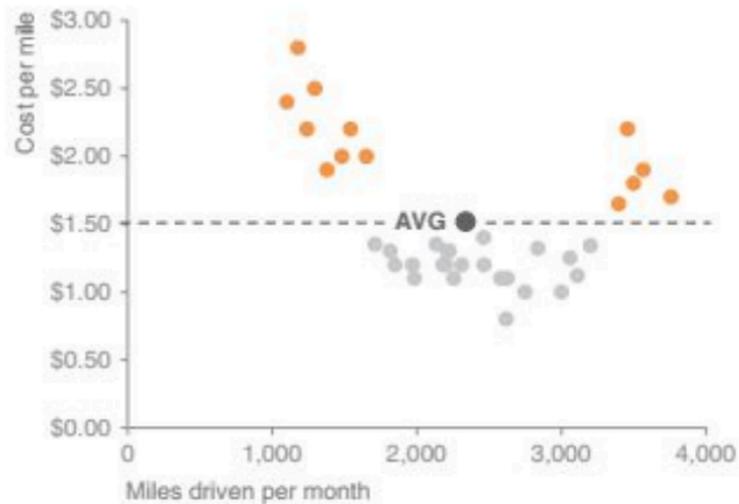


Stephen S. Brier and Stephen E. Fienberg, "Recent Econometric Modelling of Crime and Punishment: Support for the Deterrence Hypothesis?" in Stephen E. Fienberg and Albert J. Reiss, Jr., eds., *Indicators of Crime and Criminal Justice: Quantitative Studies* (Washington, D.C., 1980), p. 89.

Scatterplots

Scatterplots can be useful for showing the relationship between two things, because they allow you to encode data simultaneously on a horizontal x-axis and vertical y-axis to see whether and what relationship exists.

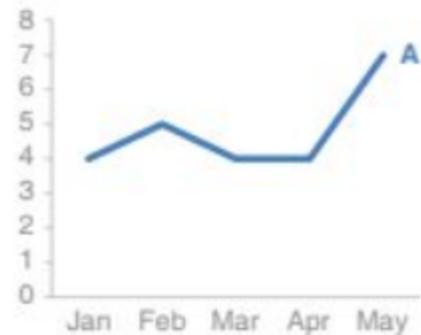
Cost per mile by miles driven



Line

Line graphs are most commonly used to plot continuous data. Because the points are physically connected via the line, it implies a connection between the points that may not make sense for categorical data (a set of data that is sorted or divided into different categories). Often, our continuous data is in some unit of time: days, months, quarters, or years.

Single series



Two series



Multiple series



Slopegraph

Slopegraphs can be useful when you have two time periods or points of comparison and want to quickly show relative increases and decreases or differences across various categories between the two data points.

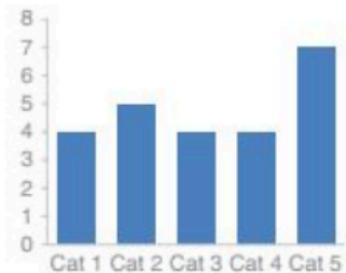
Employee feedback over time



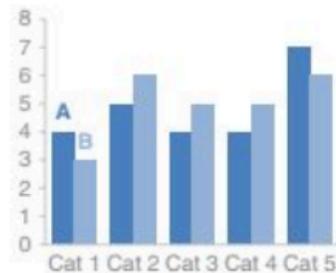
Bar

Bar charts are easy for our eyes to read. Our eyes compare the end points of the bars, so it is easy to see quickly which category is the biggest, which is the smallest, and also the incremental difference between categories. Note that, because of how our eyes compare the relative end points of the bars, it is important that bar charts always have a zero baseline

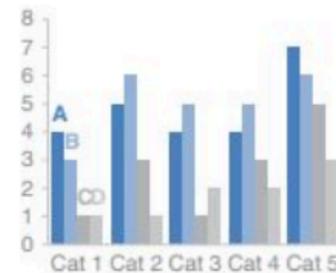
Single series



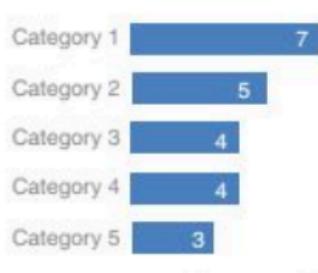
Two series



Multiple series



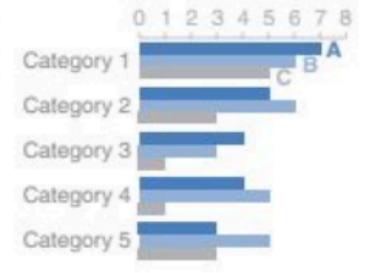
Single series



Two series



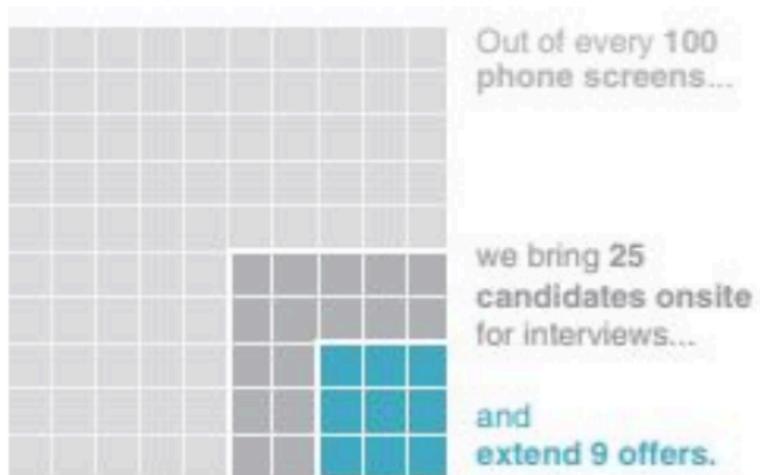
Multiple series



Area

Area graphs are to avoided except when to visualize numbers of vastly different magnitudes. The second dimension you get using a square for this allows this to be done in a more compact way than possible with a single dimension.

Interview breakdown



To be avoided

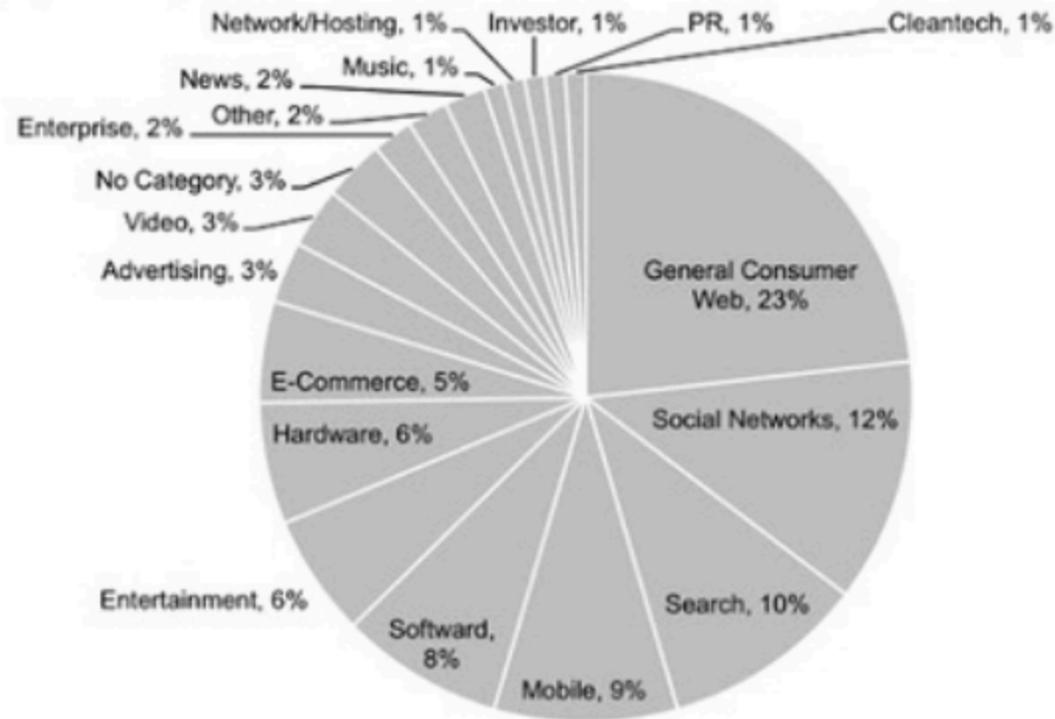
- Pie and donut charts
- 3D unless plotting in 3D
- Secondary Y-axis
- Chartjunk

Pie Charts

Pie charts are really hard for people to read!

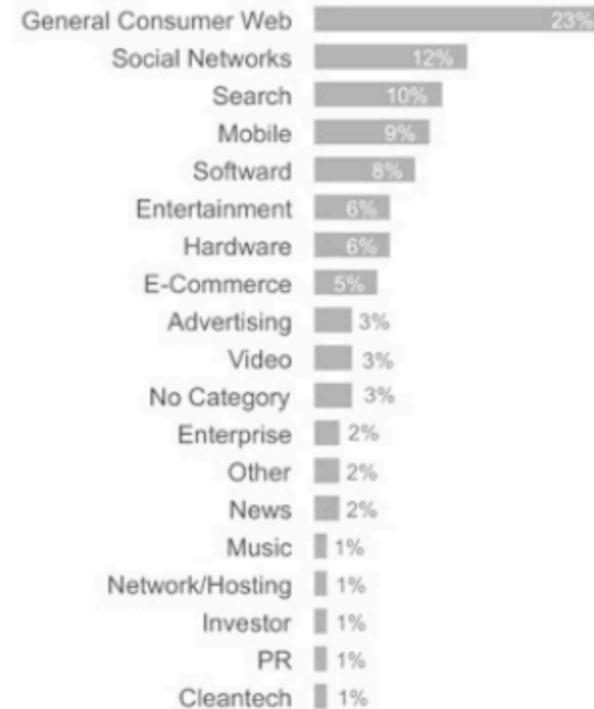
TechCrunch Coverage: 2005 - 2011

A slightly better pie?



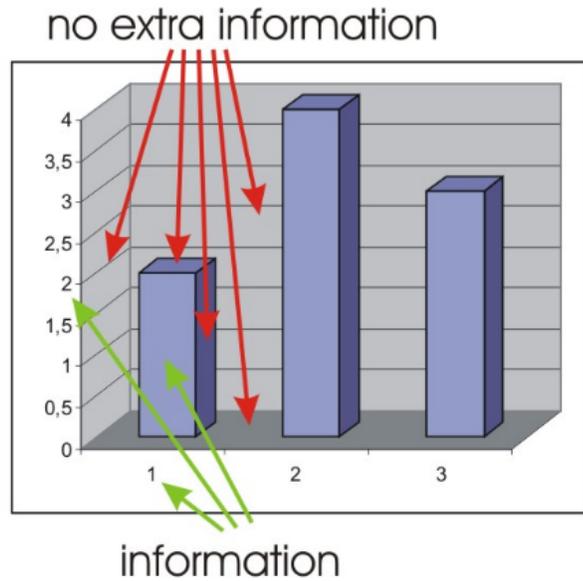
TechCrunch Coverage: 2005 - 2011

Bars are best!



3D

Never use 3D to plot a single dimension. 3D skews our numbers, making them difficult or impossible to interpret or compare. Adding 3D to graphs introduces unnecessary chart elements



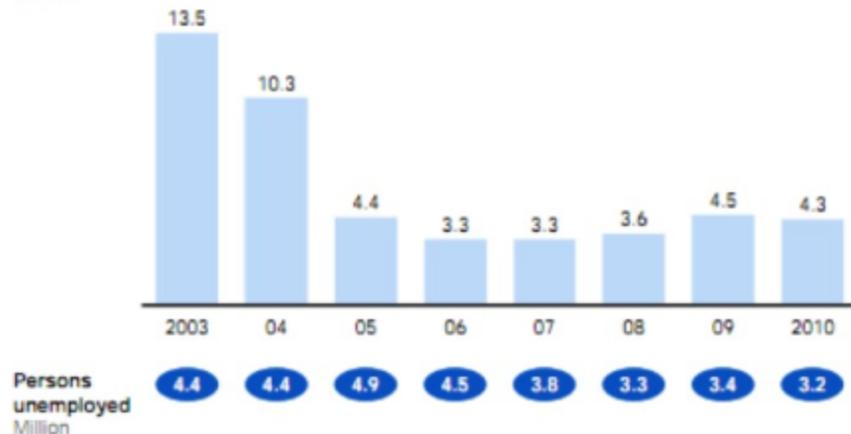
Secondary X or Y-axis

When interpreting the figure, it takes some time and reading to understand which data should be read against which axis.

Exhibit 18

Germany has achieved a significant reduction in its spending on active labor market policies without an increase in national unemployment

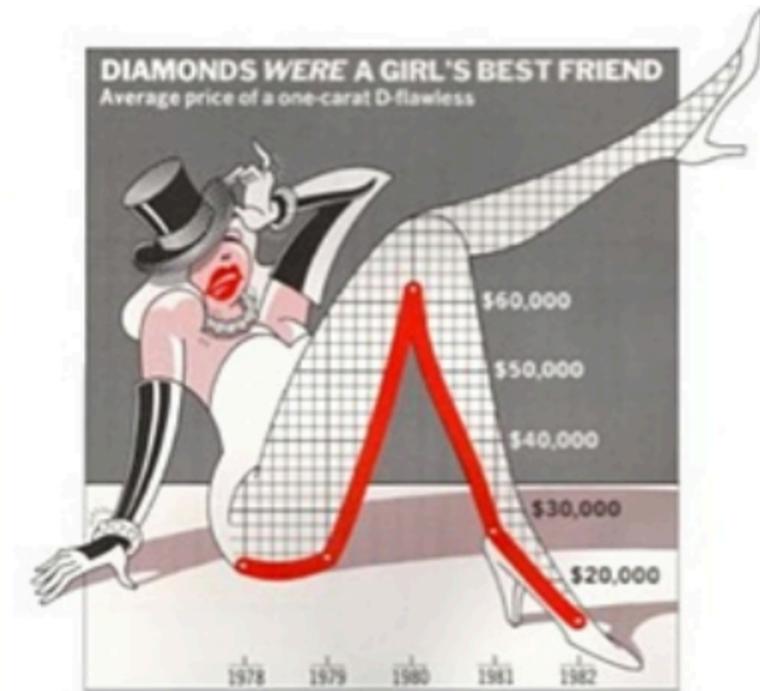
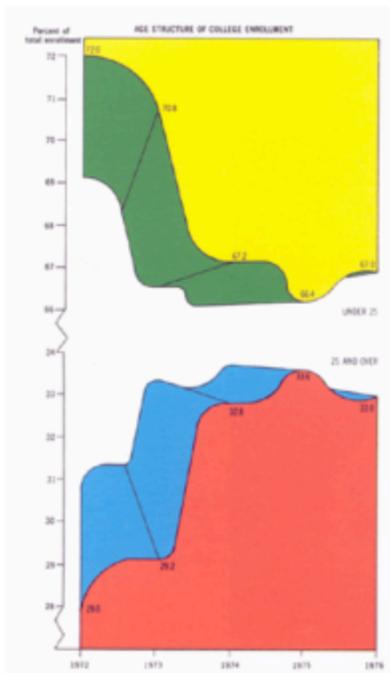
Disposable budget for active labor market policies, 2003–10
€ billion



SOURCE: Bundesagentur für Arbeit; McKinsey Global Institute analysis

Chartjunk

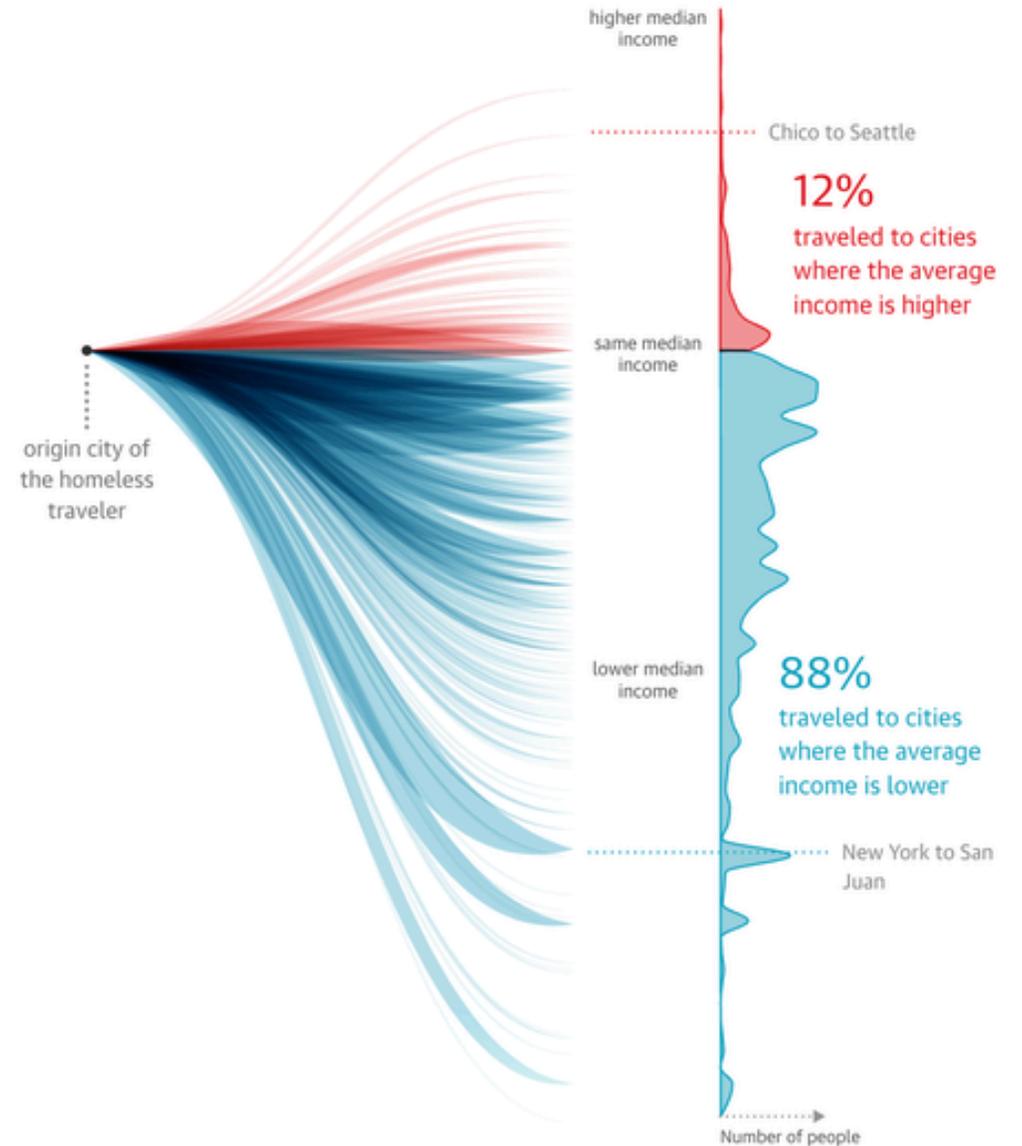
Examples of unnecessary elements that might be called chartjunk include heavy or dark grid lines, unnecessary text, inappropriately complex or gimmicky font faces, ornamented chart axes, and display frames, pictures, backgrounds or icons within data graphs, ornamental shading and unnecessary dimensions.



Good Graphs

Good graphs clearly show the important features of the data. They should always have: **title**, **labelled axes** - and a **key**

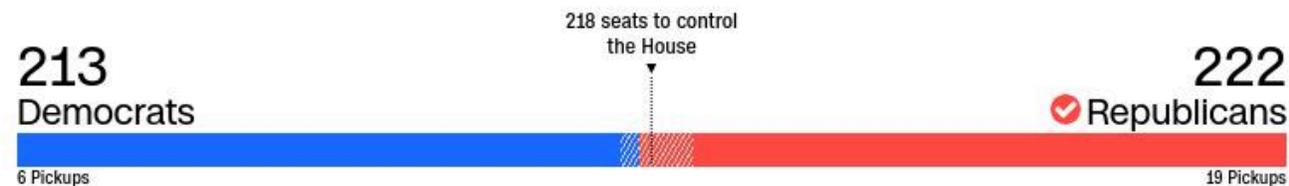
Most ticket recipients are relocated to places with a lower median income



Good graphs

They should *tell a story* - and be **memorable** - but also have a low information to ink ratio - and **not mislead** - the viewer. Choice of colour when designing charts and graphs is also important to allow for colour blindness and black and white printing.

Republicans will win the House of Representatives, CNN projects, in a victory that will fall short of their hopes of a "red wave" but thwart President Joe Biden's domestic agenda and will likely subject his White House to relentless investigations.



Rep. Donald McEachin, a Virginia Democrat, died on November 28 after being the projected winner to serve another term earlier in the month. His office will be vacant when the new Congress starts in 2023 and a special election will be held to fill the seat.

Map Options

Show Only

CNN PROJECTION

LEADING

KEY RACES

FLIPPED SEATS

Enable Scroll to Zoom

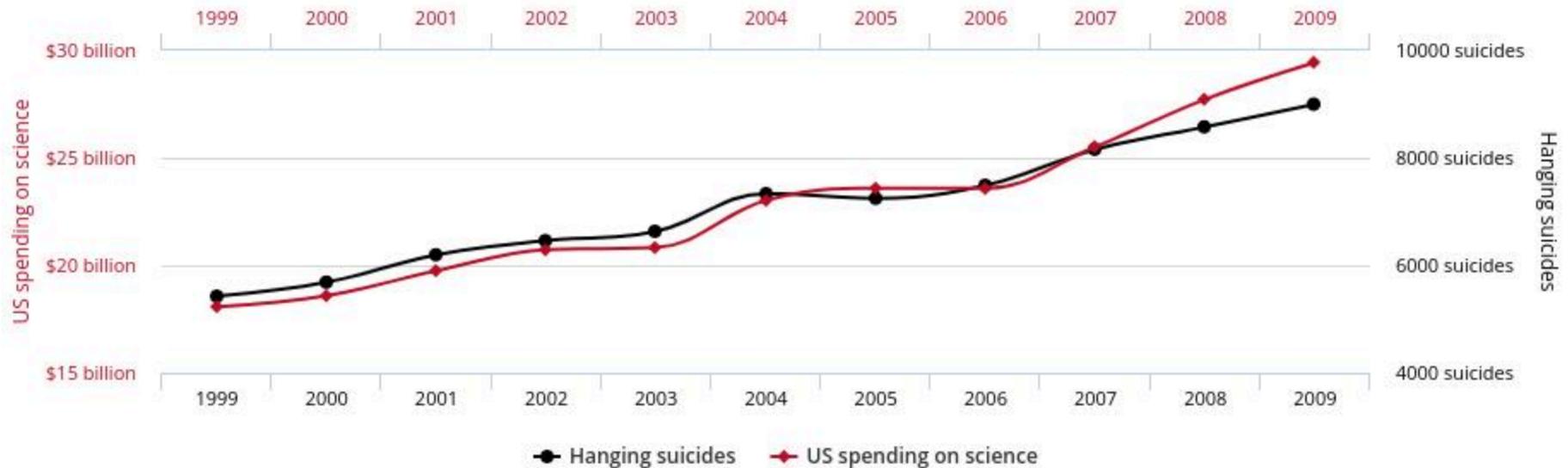


Common Mistakes

- Leaving gaps/changing the scale in vertical axes
- Uneven shading/colours
- Unfair emphasis on some sections
- Distorting areas in histograms (bar widths should always be equal)
- Use of 3-dimensions instead of two
- Misleading use of pictograms

US spending on science, space, and technology correlates with Suicides by hanging, strangulation and suffocation

Correlation: 99.79% (r=0.99789126)



tylervigen.com

Data sources: U.S. Office of Management and Budget and Centers for Disease Control & Prevention

Source: <http://www.tylervigen.com/spurious-correlations>

Books

- Edward R. Tufte. The Visual Display of Quantitative Information. Graphics 2001.
https://www.goodreads.com/book/show/17744.The_Visual_Display_of_Quantitative_Information?ac=1from_search=true&qid=cMoP2LwLen&rank=1
- Cole Nussbaumer Knaflic. Storytelling with Data: A Data Visualization Guide for Business Professionals. Wiley 2015.
<https://www.goodreads.com/book/show/26535513-storytelling-with-data>

Links

- The Few, The Proud: 11 Key Principles of Effective Data Visualization
<https://www.business2community.com/big-data/the-few-the-proud-11-key-principles-of-effective-data-visualization-02076890>
- Data Visualization Hacks <https://uxdesign.cc/data-visualization-hacks-75d56d5bfa66>
- Data Visualization UX Best Practices (Updated 2024).
<https://www.designstudiouiux.com/blog/data-visualization-ux-best-practices/>
- Line Chart Design Made Simple. <https://uxdesign.cc/line-chart-design-made-simple-a1b823510674>

Links

- Designing Charts: principles every designer should know (part 2)
<https://uxdesign.cc/designing-charts-principles-every-designer-should-know-part-2-ce1e06af56fc>
- Building Blocks of Good Graphs. <https://www.chaione.com/blog/building-blocks-graphs>
- Data Visualization Design, Part 4: Removing Chart Junk.
<https://medium.com/@LauraHKahn/data-visualization-design-part-4-removing-chart-junk-28b3bdd0faa1>
- COVID-19 In Charts: Examples of Good & Bad Data Visualization.
<https://analytical.com/blog/covid19-in-charts>

Thank you!