SCIENTIFIC DATA VISUALIZATION

AGENDA

Introduction

Basics of data visualization

Basic guidelines and common pitfalls

☕ **Break**

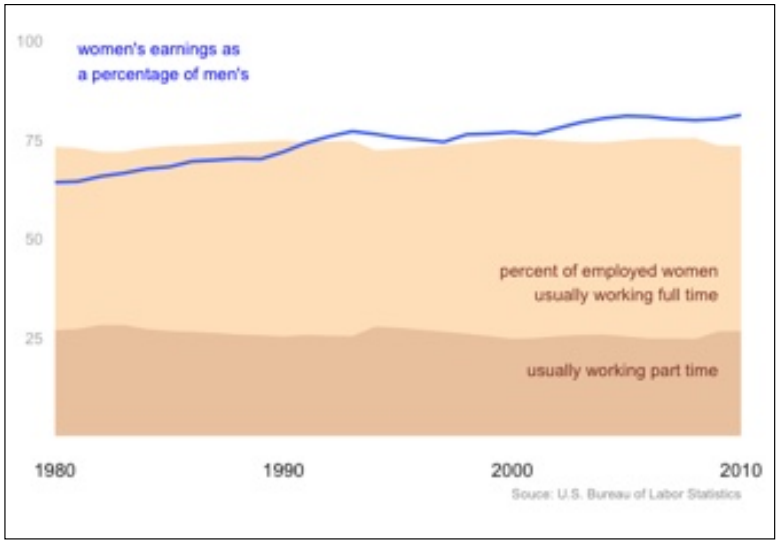
Examples and exercises in ggplot2

GEORGIOS KARAMANIS

Psychiatrist, Gender identity clinic, Uppsala University Hospital

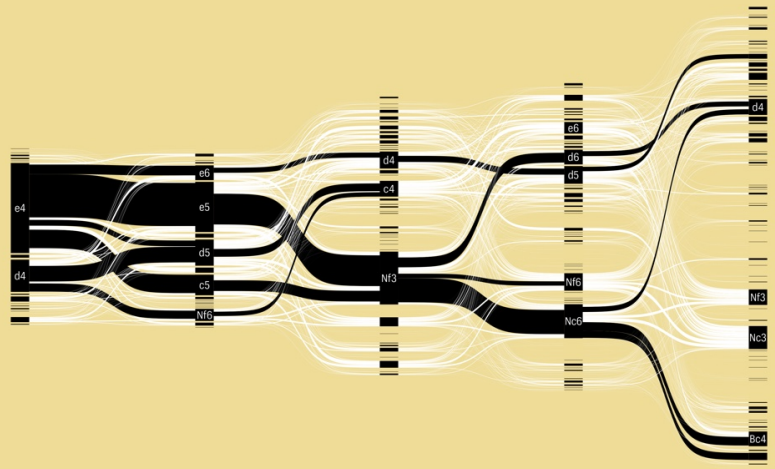
PhD student, Institution for medical sciences, Uppsala University

Data visualization designer, Explained

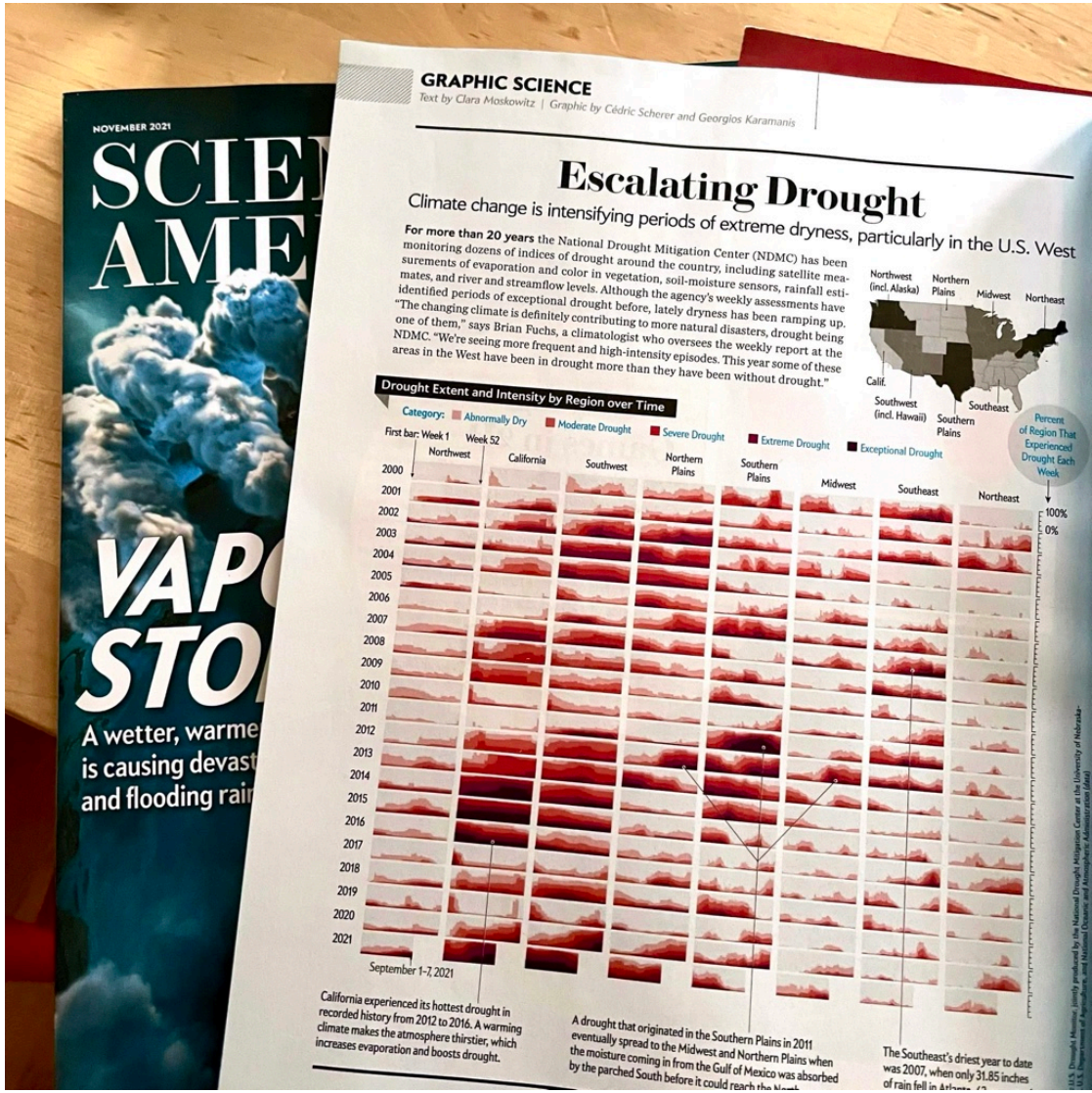


First Five Moves

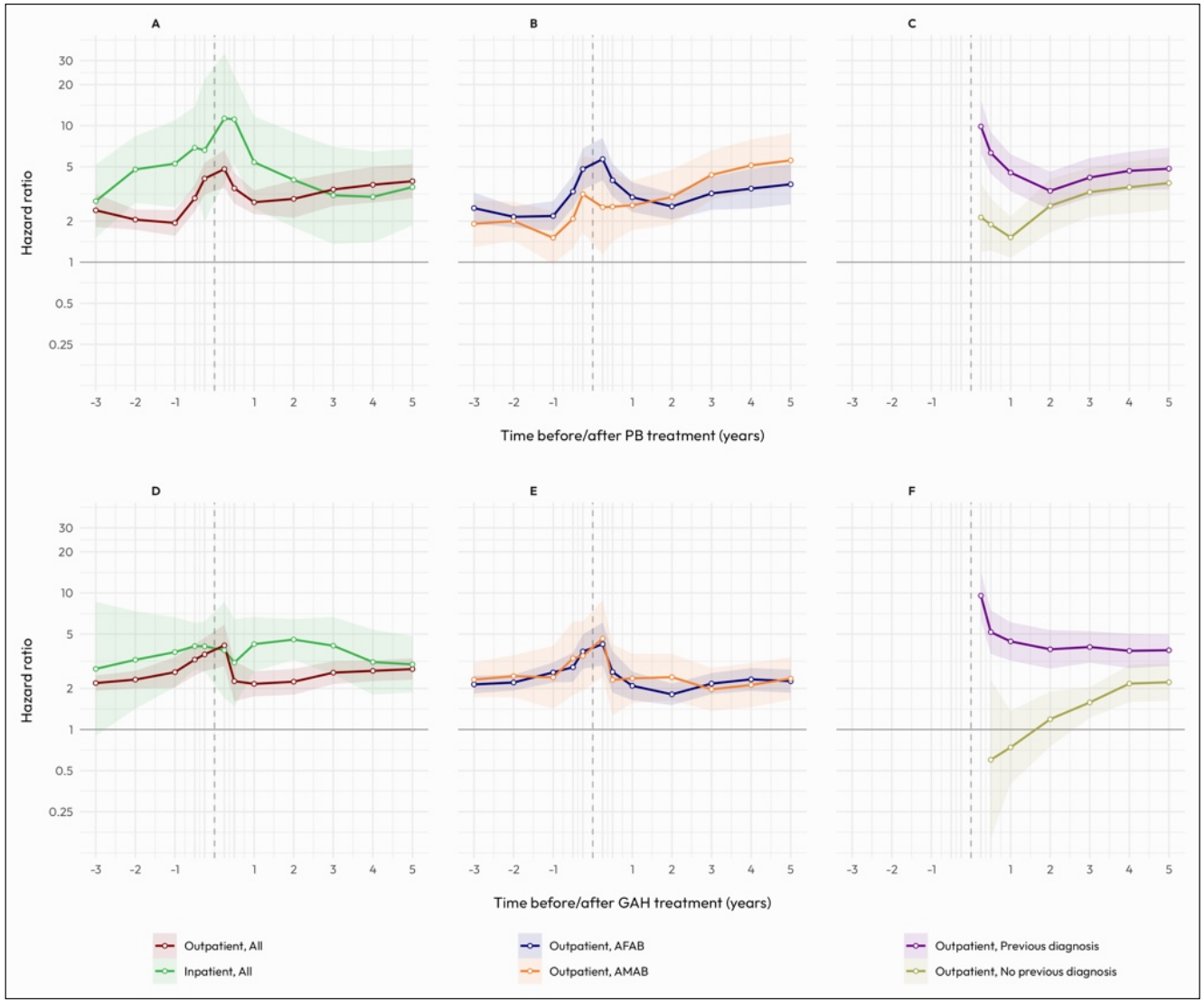
Flow of opening moves in over 20 000 chess games from Lichess.org. The most popular moves are highlighted in black, revealing common opening strategies.

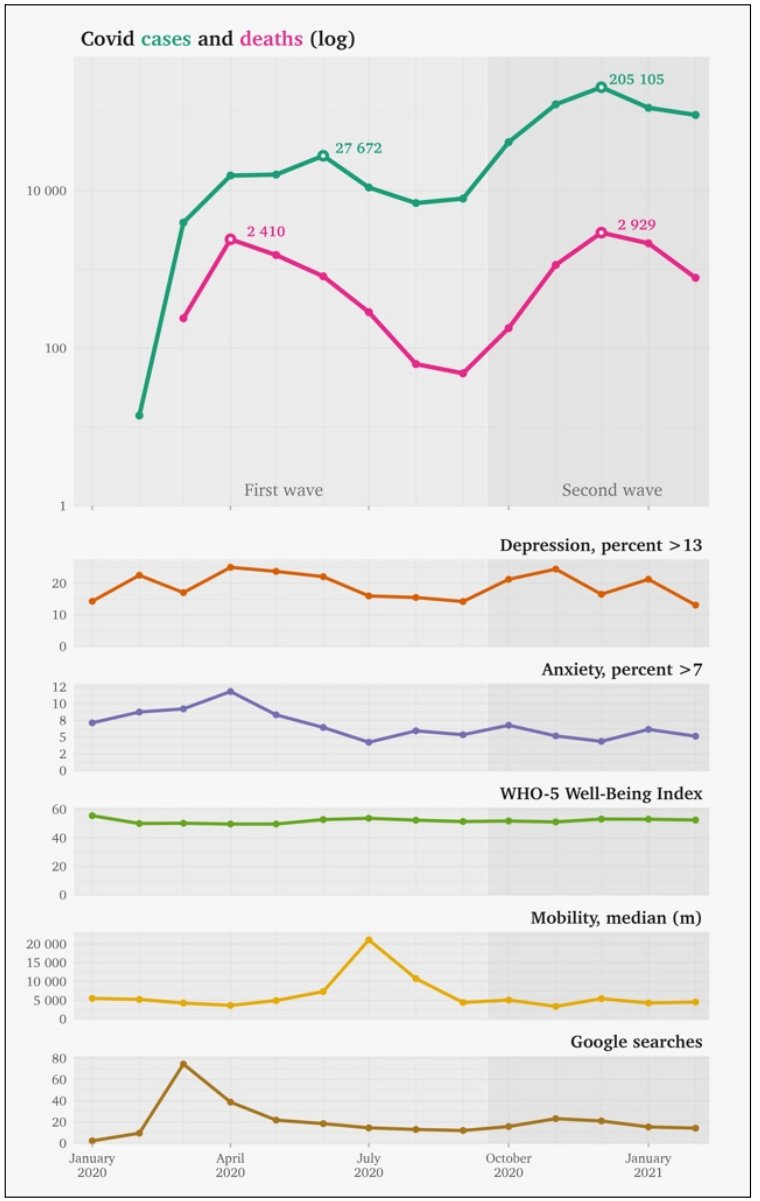


Source: Lichess.org via Kaggle/Mitchell J. · Graphic: Georgios Karamanis



https://www.scientificamerican.com/article/climate-change-drives-escalating-drought/#

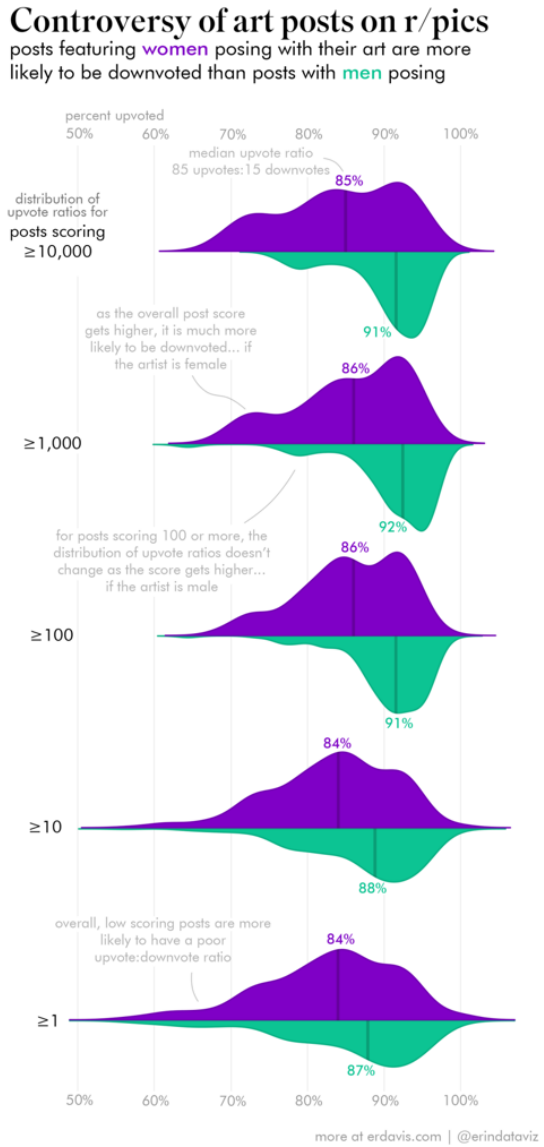

A
B
C
D
E
F
Hazard ratio
Time before/after PB treatment (years)
Time before/after GAH treatment (years)
Outpatient, All
Inpatient, All
Outpatient, AFAB
Outpatient, AMAB
Outpatient, Previous diagnosis
Outpatient, No previous diagnosis


Covid cases and deaths (log)
205 105
27 672
2 410
2 929
First wave
Second wave
Depression, percent >13
Anxiety, percent >7
WHO-5 Well-Being Index
Mobility, median (m)
Google searches
January 2020
April 2020
July 2020
October 2020
January 2021

INTRODUCTION

DATA VISUALIZATION

Data visualization blends art and science to convey stories from data via graphical representations.

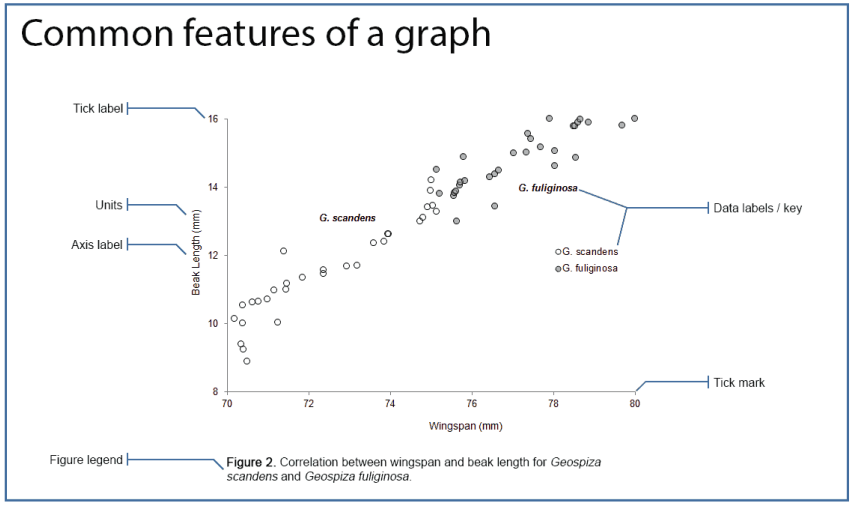


https://erdavis.com/2021/06/14/do-women-who-pose-with-their-art-on-reddit-get-more-upvotes/

SCIENTIFIC CHARTS

Figures in scientific publications are critically important because they often show the data supporting key findings

Weissgerber TL, Milic NM, Winham SJ, Garovic VD (2015) Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLoS Biol 13(4): e1002128. https://doi.org/10.1371/journal.pbio.1002128



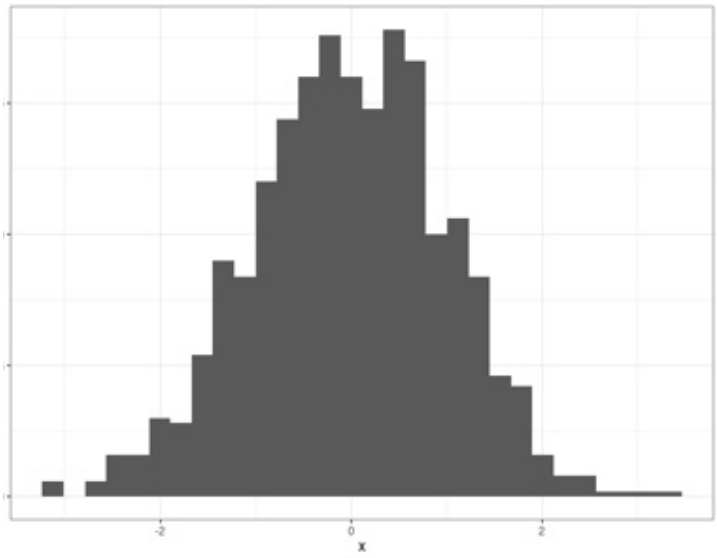
https://www.clips.edu.au/displaying-data/

WHY CHARTS?

Easier overview of data

(especially complex)

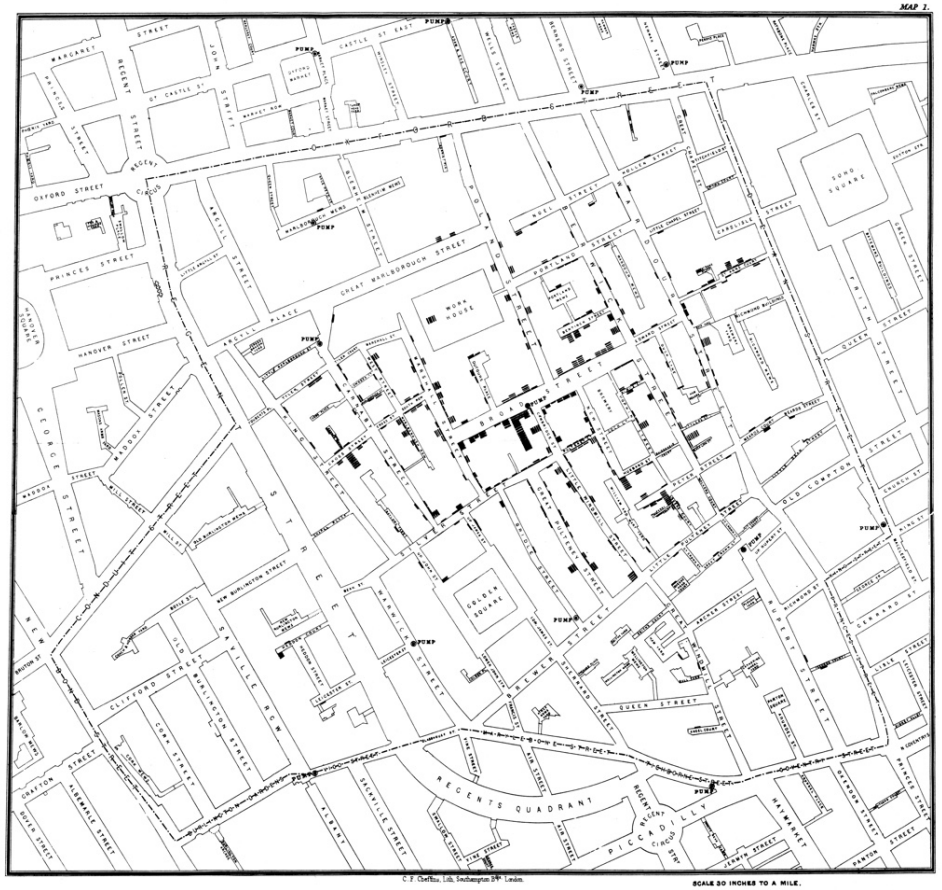
	x
1	0.376935655
2	-0.603182135
3	0.659734512
4	-1.597222895
5	-0.645683665
6	-0.929817315
7	1.401684999
8	0.520906260
9	0.369003441
10	0.573936602
11	1.304568093
12	-2.537849064
13	0.920903182
14	1.419055702
15	-0.288114527
16	1.842139177
17	0.647224489
18	1.793445537
19	-2.002214213
20	0.885130196
21	-0.369039857
22	-0.136582806



WHY CHARTS?

Find new patterns

BROAD STREET PUMP OUTBREAK, JOHN SNOW, 1854

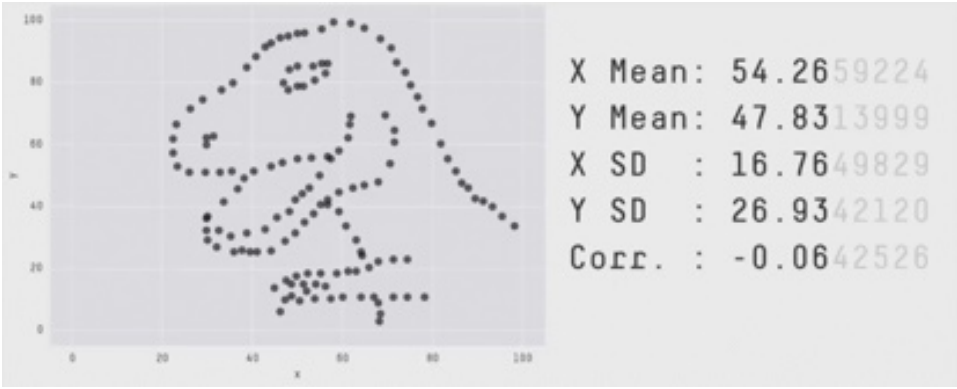


https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

WHY CHARTS?

Easier for comparisons of data and their statistics

THE DATASAURUS DOZEN

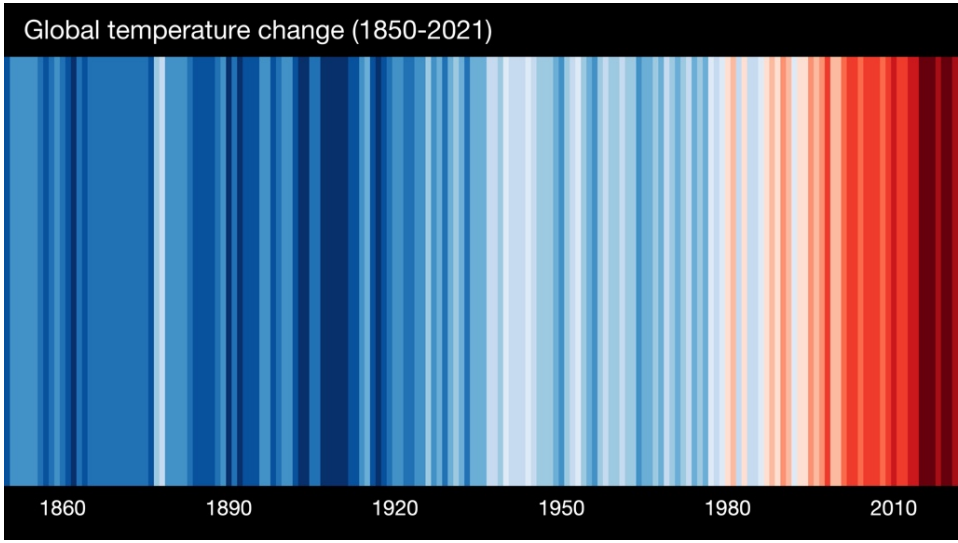


https://www.autodesk.com/research/publications/same-stats-different-graphs

WHY CHARTS?

To better communicate your point

WARMING STRIPES, ED HAWKINGS, 2018

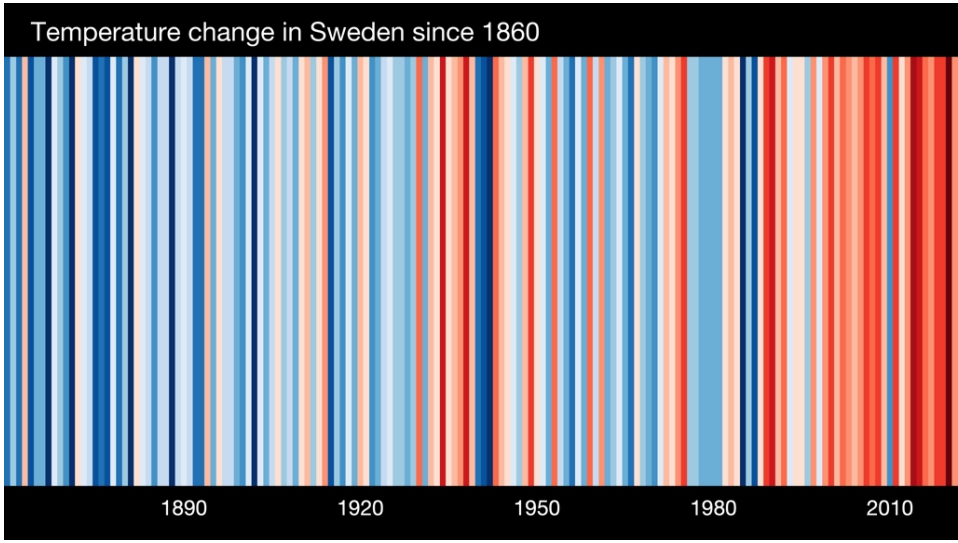


https://www.climate-lab-book.ac.uk/2018/warming-stripes/

WHY CHARTS?

To better communicate your point

WARMING STRIPES, ED HAWKINGS, 2018

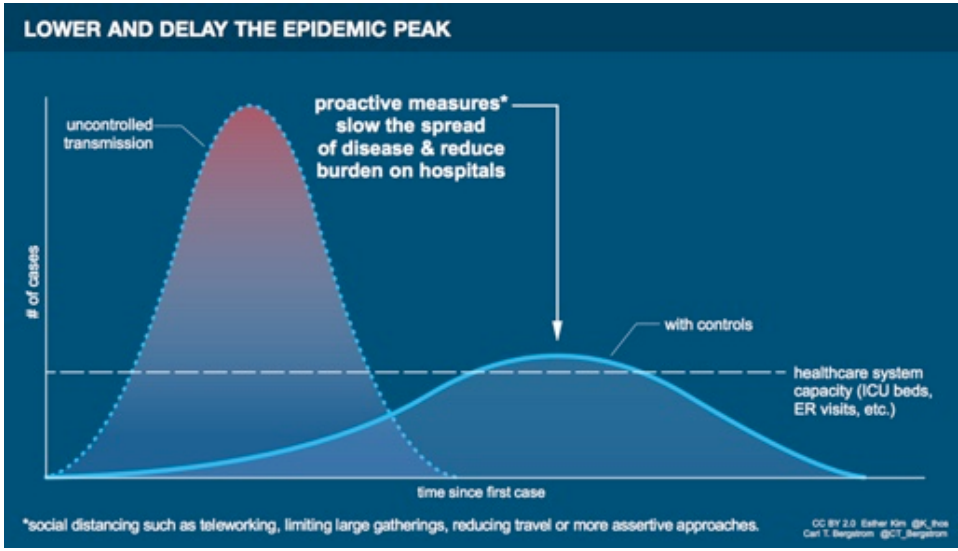


https://showyourstripes.info/l/europe/sweden/all

WHY CHARTS?

They look good, are memorable and shareable

ESTHER KIM & CARL T. BERGSTROM, 2019



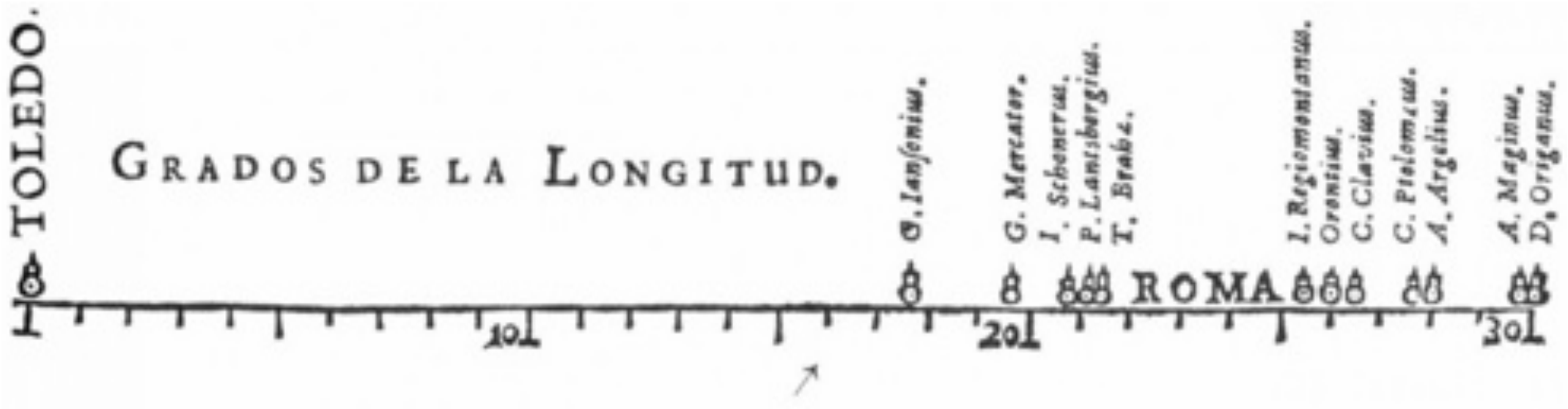
http://ctbergstrom.com/covid19.html


we're sick of this SHIT
1850
2018


MUN 0 - 0 REA 28:04

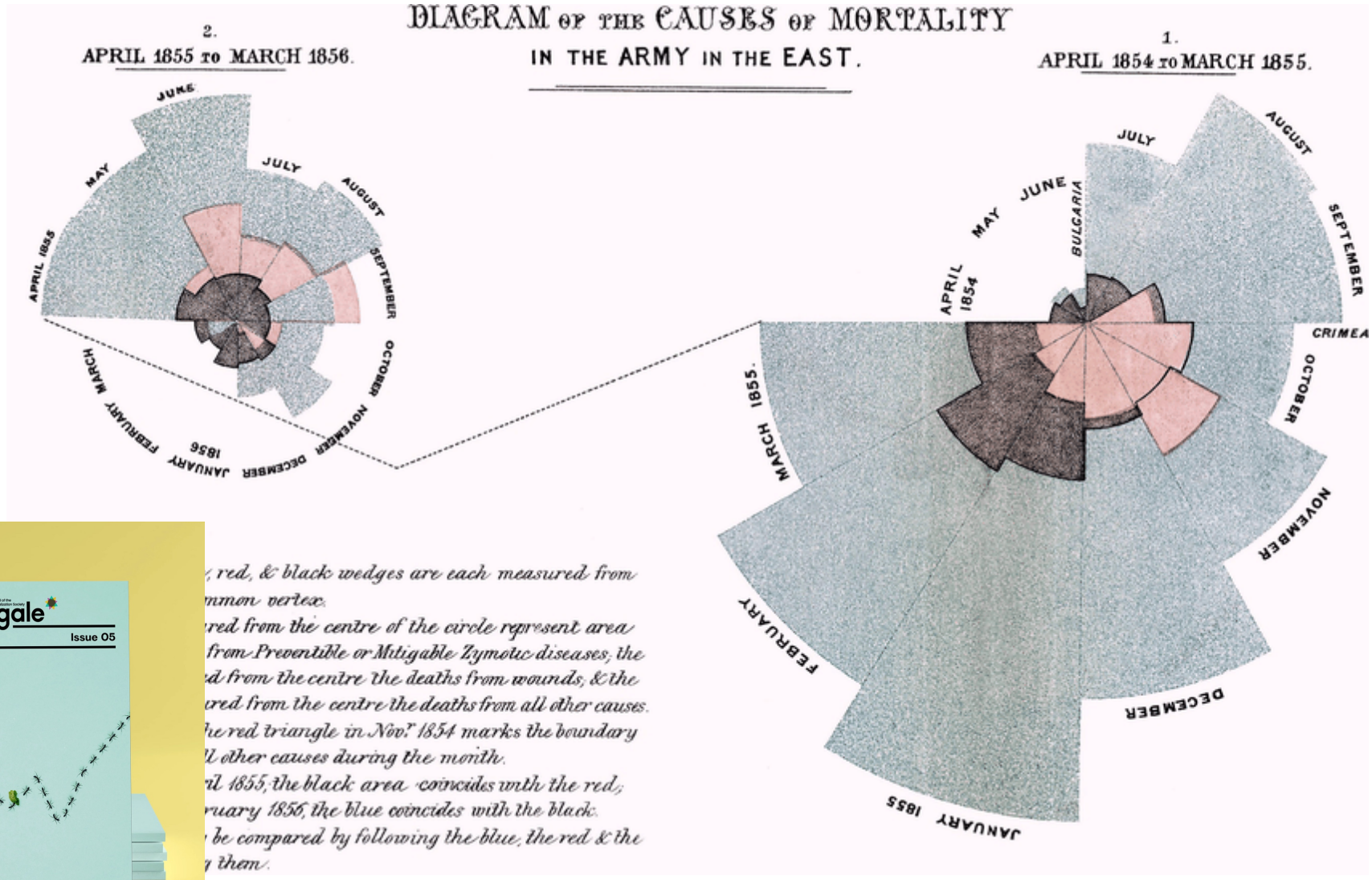
SCIENTIFIC CHARTS HAVE A LONG HISTORY

MICHAEL FLORENT VAN LANGREN, 1644



https://historyofinformation.com/detail.php?id=3415

DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST, FLORENCE NIGHTINGALE, 1858



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

BAD CHARTS ARE HARD TO READ, MUDDLE OR HARM OUR MESSAGE AND EVEN OUR CREDIBILITY

~~CLINICAL~~ → MEANINGFUL ← ~~BEAUTIFUL~~

John Burn-Murdoch
Making meaningful graphics

BASICS OF DATA VISUALIZATION

VISUAL ELEMENTS

position

length

angle

direction

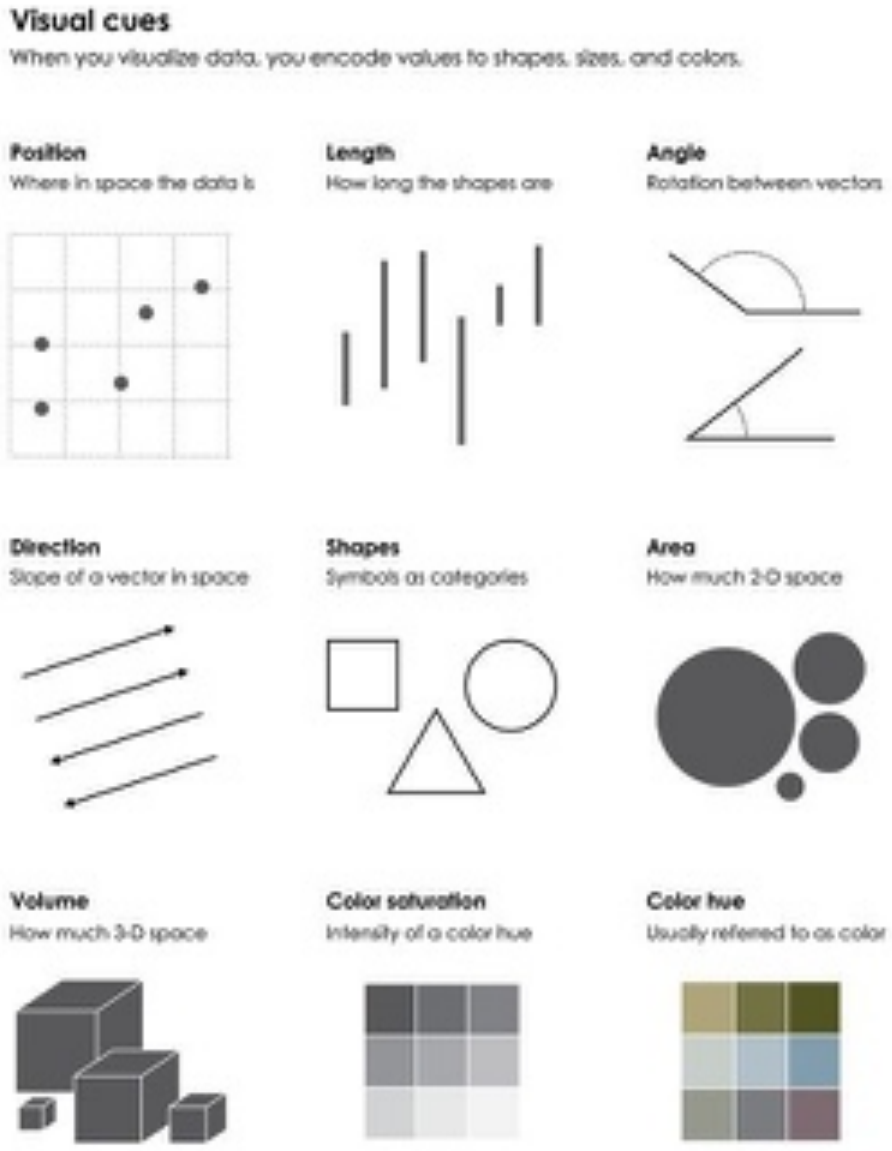
shape

area

volume

saturation/lightness

color hue

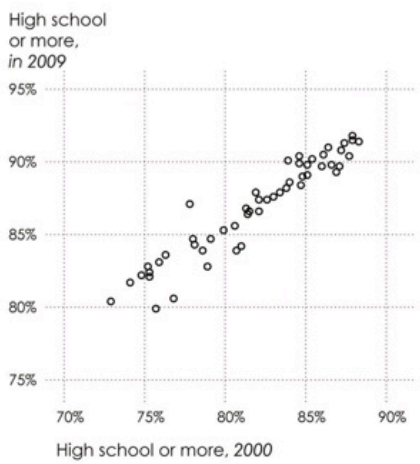


https://flowingdata.com/data-points/DataPoints-Ch3.pdf

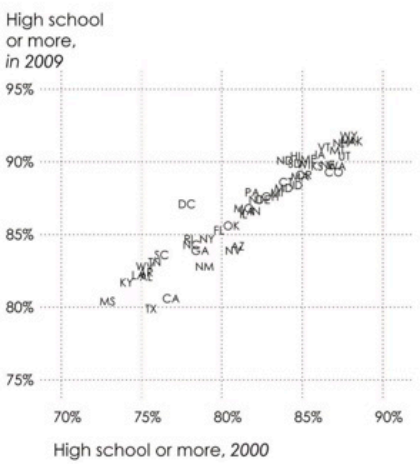
VISUAL ELEMENTS

combinations

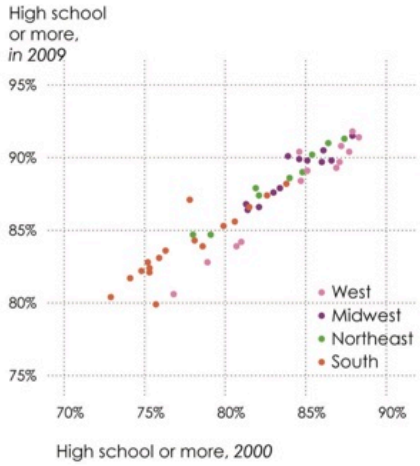
Position



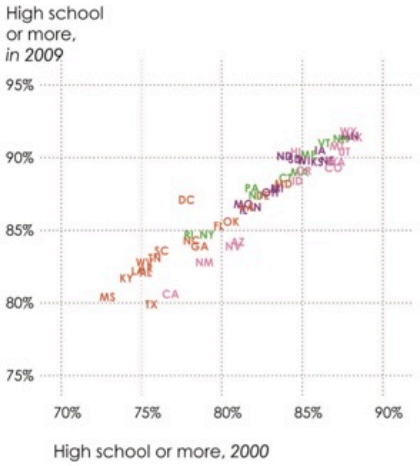
Position + Symbols



Position + Color

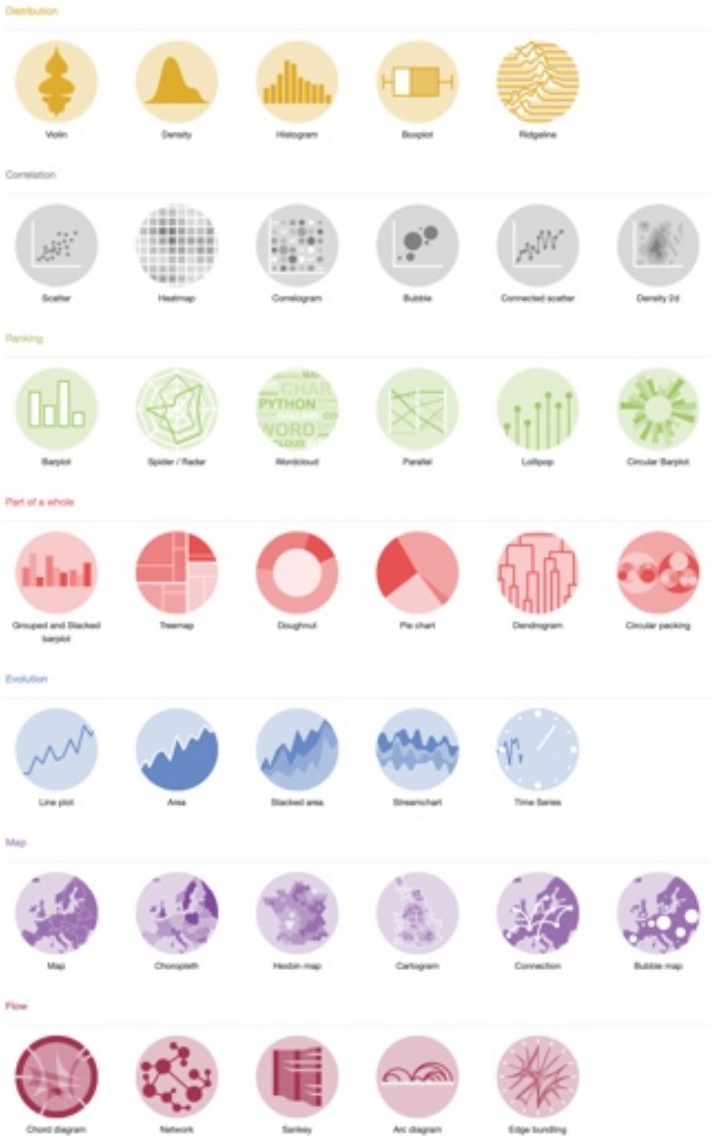


Position + Symbols + Color



https://flowingdata.com/data-points/DataPoints-Ch3.pdf

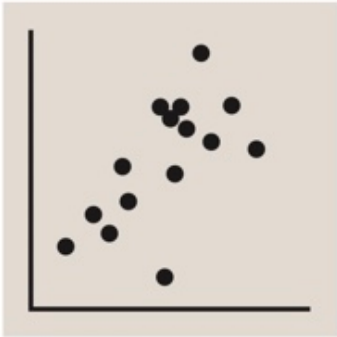
BASIC TYPES OF CHARTS



https://r-graph-gallery.com

SCATTERPLOT

Scatterplot



The standard way to show the relationship between two continuous variables, each of which has its own axis.

LINE CHART

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

HISTOGRAM

Histogram

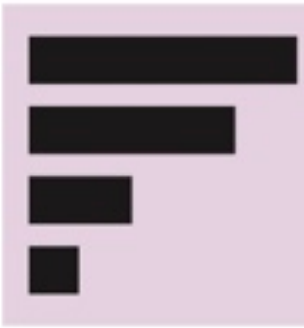
The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

BAR CHART

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar



See above. Good when the data are not time series and labels have long category names.

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation and unemployment, income and life expectancy.

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Column + line timeline

A good way of showing the relationship between an amount (column(s)) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less effective at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (ages/sex) distribution, revealing inequality

Histogram

The standard way to show a statistical distribution - keep the gaps between columns small to highlight the 'shape' of the data.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution, effectively, back to back histograms.

Cumulative curve

A good way of showing how unequal a distribution is/ y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Beeswarm

Use to emphasise individual points in a distribution. Points can be sized to an additional variable. Best with medium-sized datasets

Change over Time

Give emphasis to changing trends. These can be short (mins-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series, sectoral changes in a market

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time - but usually best with only one series of data at a time.

Column + line timeline

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes in total, but very large or components can be very difficult.

Candlestick

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priorities timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Streamgraph

A type of area chart, use when seeing changes in proportions over time is more important than individual values

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation, volumes in general

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an axis to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space efficient way of showing value of multiple variables – but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column/bar

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising parliamentary composition by number of seats.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in small multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values.

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity-colour scale. As choropleth map – but not wrapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information, relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other: good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

Visual Vocabulary

Read the FT for free

Add fresh examples and insights to your work, make more informed career choices, and feel more confident managing money with the Financial Times.

Teachers can register your school and students can create a login by scanning here.

ft.com/schoolsarefree

FT graphic: Alan Smith, Chris Campbell, Ian Bott, Liz Faunce, Graham Parrish, Billy Ehrenberg-Shannon, Paul McLellan, Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

FT FINANCIAL TIMES

Royal Geographical Society with IBG
Advancing geography and geographical learning

BASIC GUIDELINES AND COMMON PITFALLS

Ten guidelines for effective data visualization in scientific publications

Environmental Modelling & Software
Volume 26, Issue 6, June 2011, Pages 822-827

Short communication

Ten guidelines for effective data visualization in scientific publications

Christa Kelleher, Thorsten Wagener

Show more

Share Cite

https://doi.org/10.1016/j.envsoft.2010.12.006

Get rights and content

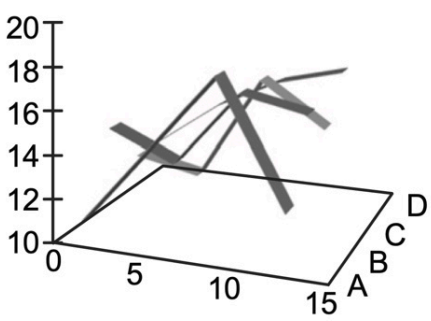
Abstract

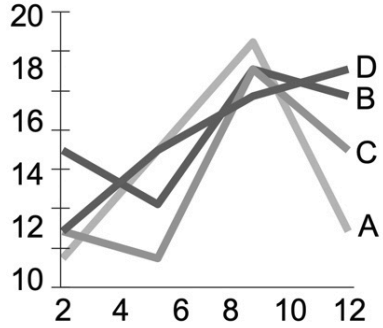
Our ability to visualize scientific data has evolved significantly over the last 40 years. However, this advancement does not necessarily alleviate many common pitfalls in visualization for scientific journals, which can inhibit the ability of readers to effectively understand the information presented. To address this issue within the context of visualizing environmental data, we list ten guidelines for effective data visualization in scientific publications. These guidelines support the primary objective of data visualization, i.e. to effectively convey information. We believe that this small set of guidelines based on a review of key visualization literature can help researchers improve the communication of their results using effective visualization. Enhancement of environmental data visualization will further improve research presentation and communication within and across disciplines.

Christa Kelleher, Thorsten Wagener, Ten guidelines for effective data visualization in scientific publications, Environmental Modelling & Software, Volume 26, Issue 6, 2011, Pages 822-827, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2010.12.006.

GUIDELINE 1

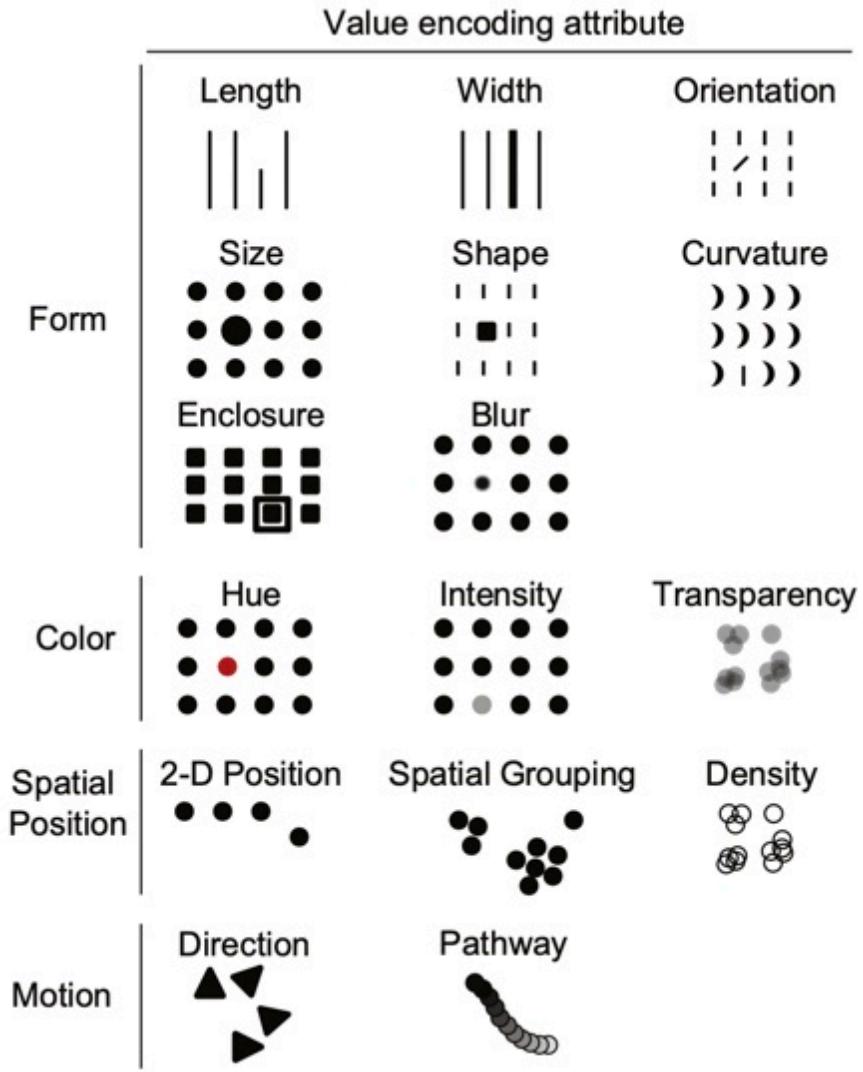
Create the simplest graph that conveys the information you want to convey





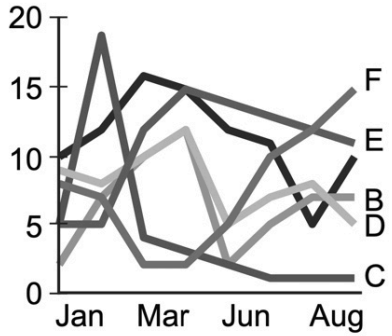
GUIDELINE 2

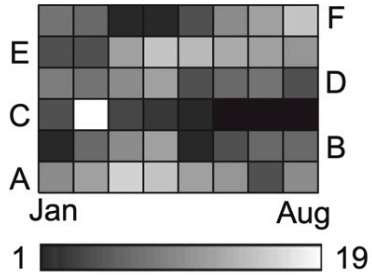
Consider the type of encoding object and attribute used to create a plot



GUIDELINE 3

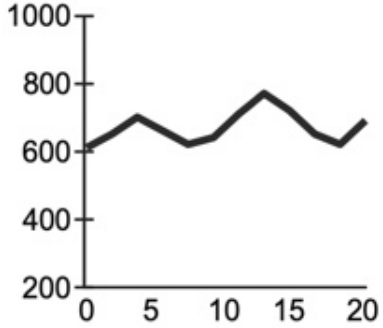
Focus on visualizing patterns or on visualizing details, depending on the purpose of the plot

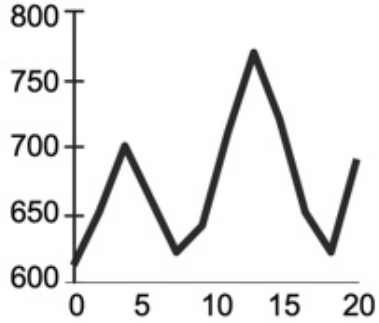




GUIDELINE 4

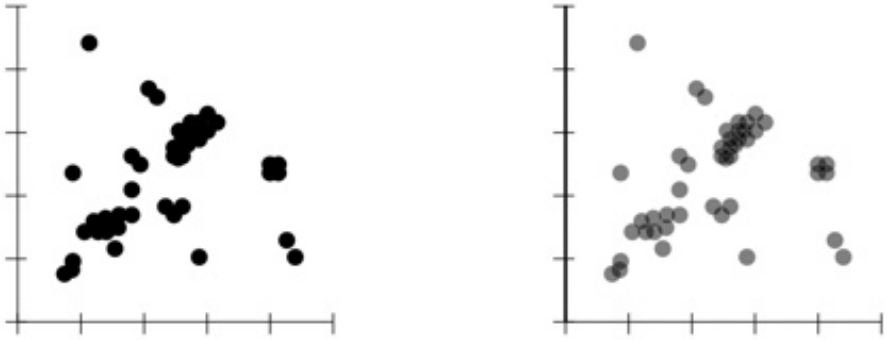
Select meaningful axis ranges





GUIDELINE 6

Plot overlapping points in a way that density differences become apparent in scatter plots

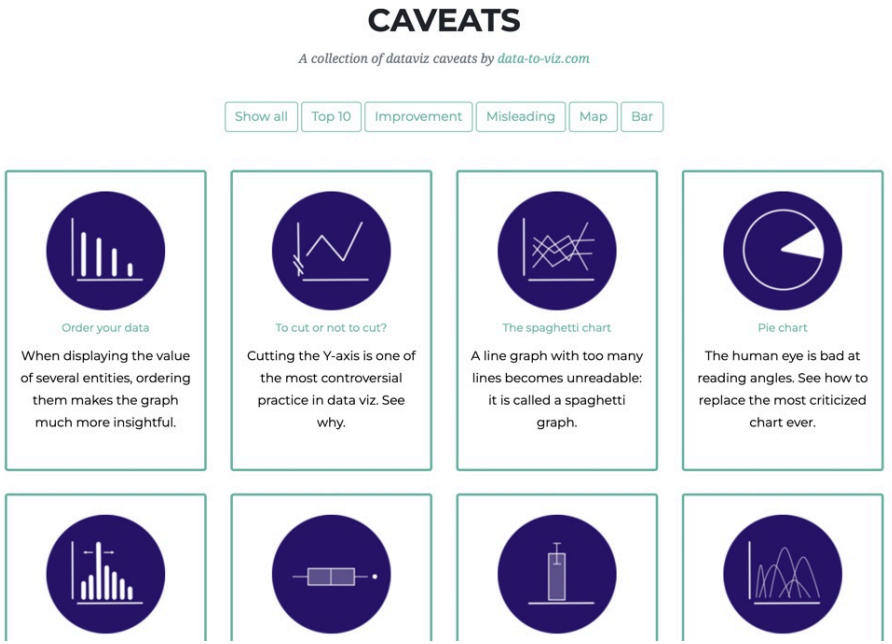


COMMON PITFALLS

and their alternatives

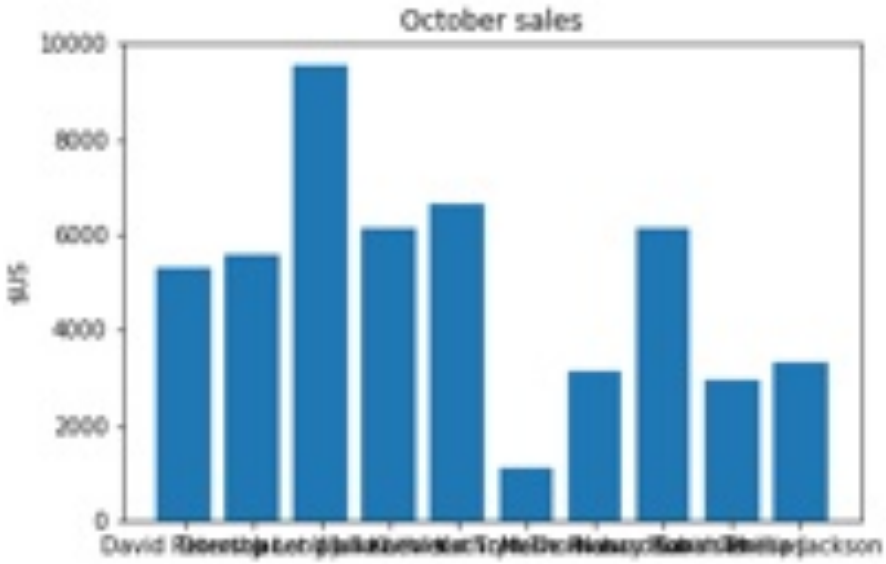
More at

https://www.data-to-viz.com/caveats.html



VERTICAL AXIS TEXT

Common problem

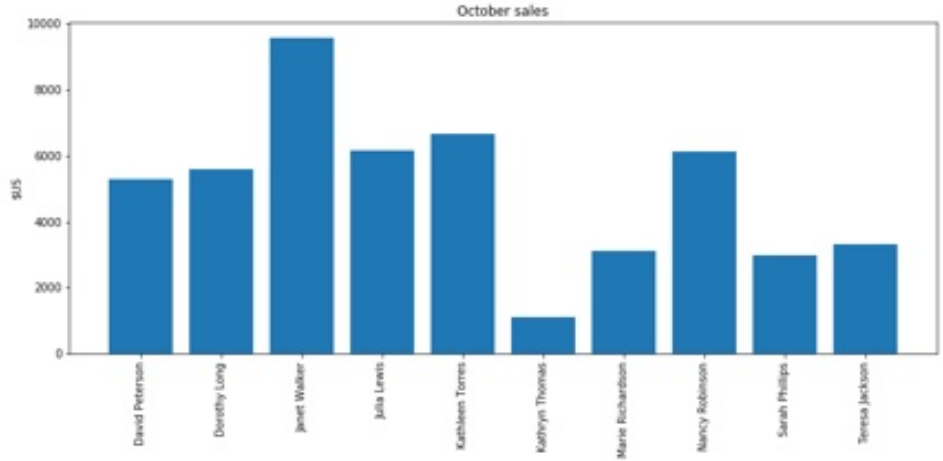


https://gorelik.net/2017/11/23/how-to-make-a-graph-less-readable-rotate-the-text-labels/

VERTICAL AXIS TEXT

Solved?

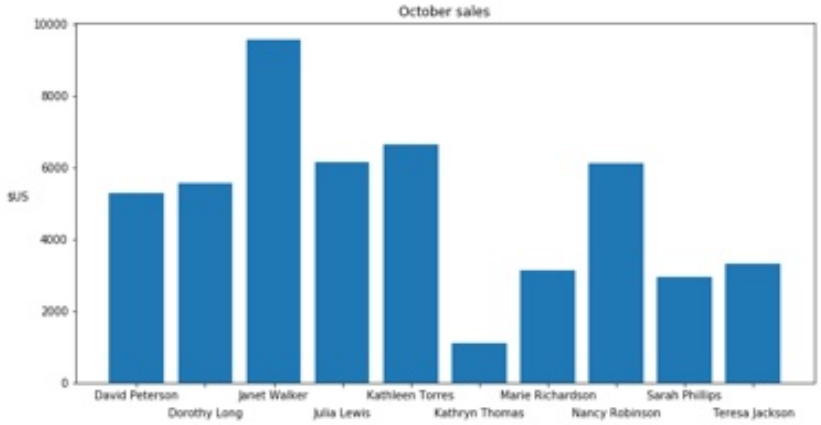
Difficult to read



https://gorelik.net/2017/11/23/how-to-make-a-graph-less-readable-rotate-the-text-labels/

VERTICAL AXIS TEXT

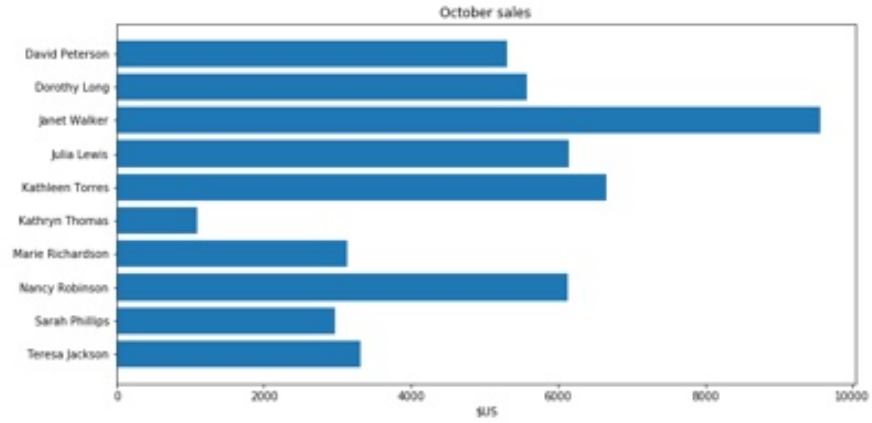
Consider moving the labels



https://gorelik.net/2017/11/23/how-to-make-a-graph-less-readable-rotate-the-text-labels/

VERTICAL AXIS TEXT

Rotate the bars

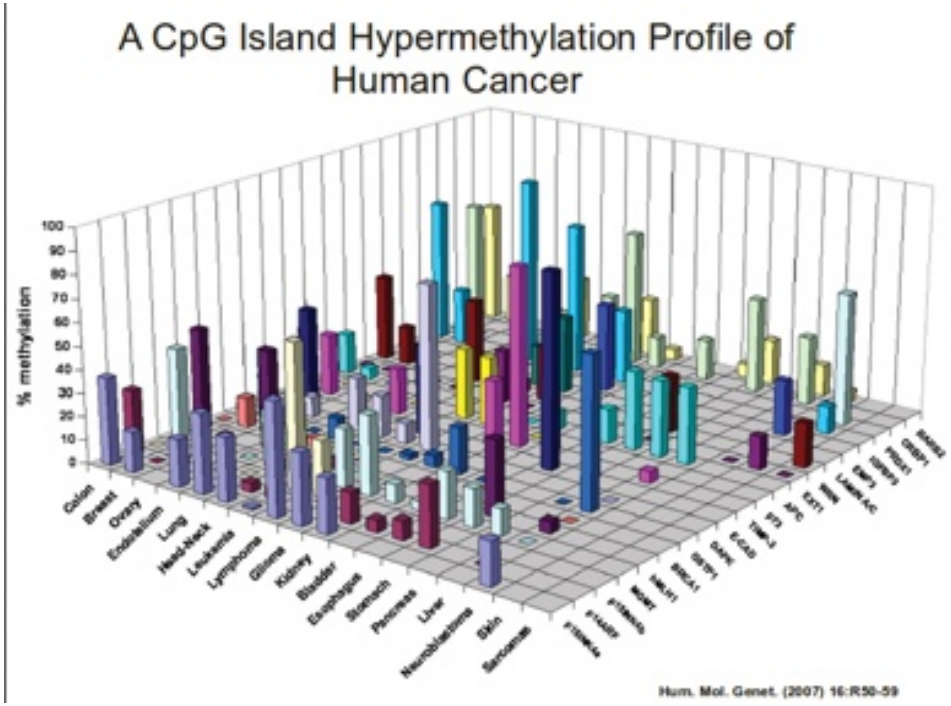


https://gorelik.net/2017/11/23/how-to-make-a-graph-less-readable-rotate-the-text-labels/

3D

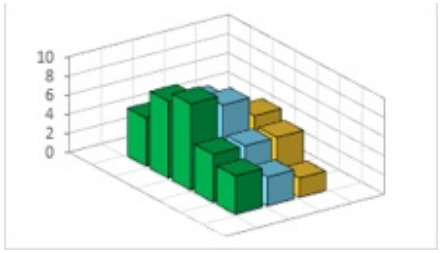
Avoid

Very difficult to read and compare, hidden data

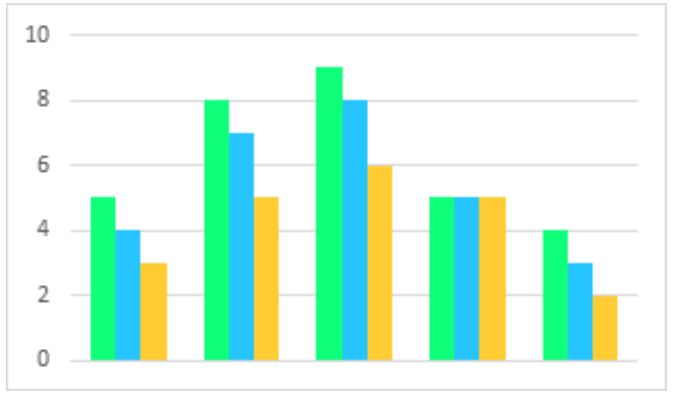


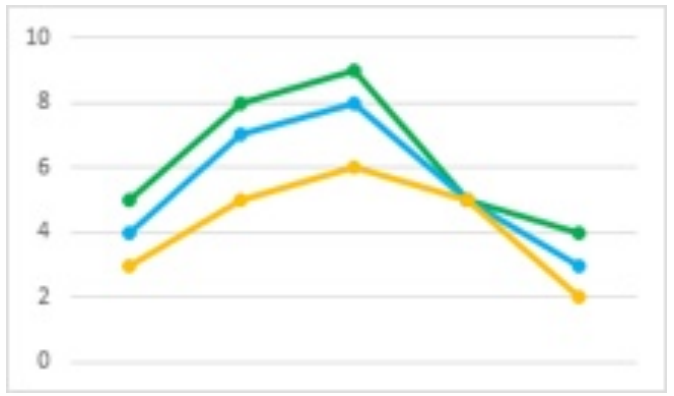
3D

Consider other chart types



https://peltiertech.com/3d-bar-chart-alternatives/





PIE CHARTS

Avoid

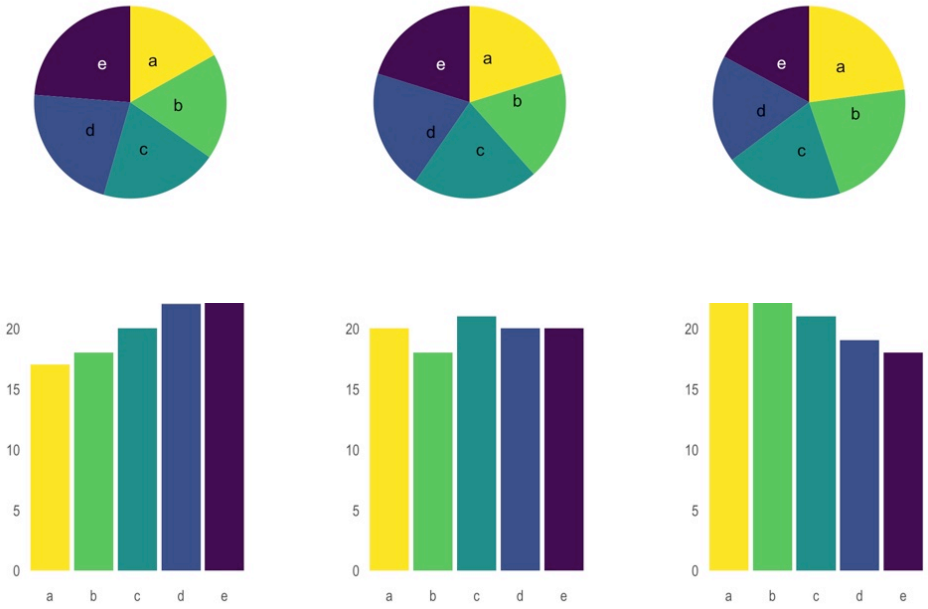
Perception problems (but not always)

OK for few categories

Can be difficult to read

PIE CHARTS

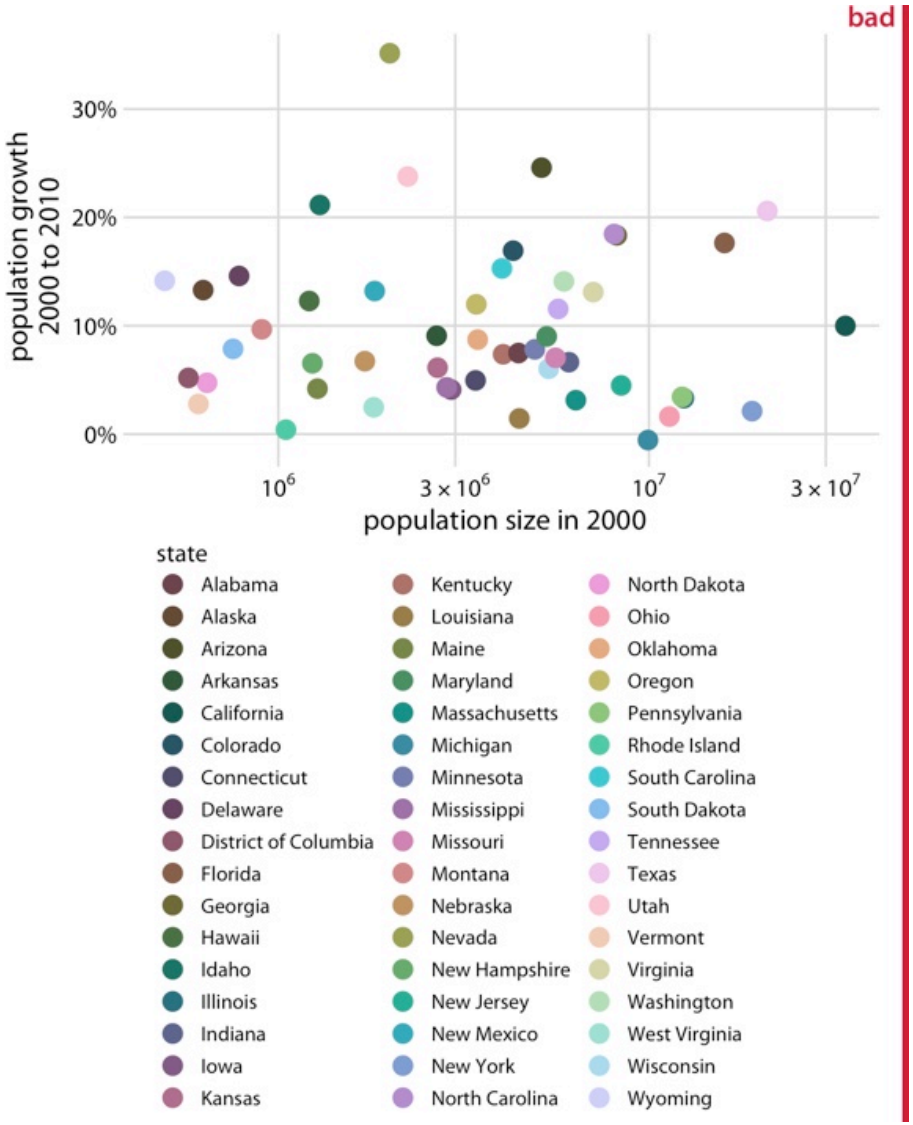
Alternative: bar chart



https://www.data-to-viz.com/caveat/pie.html

TOO MANY COLORS

OK to have less than 6-8 categories



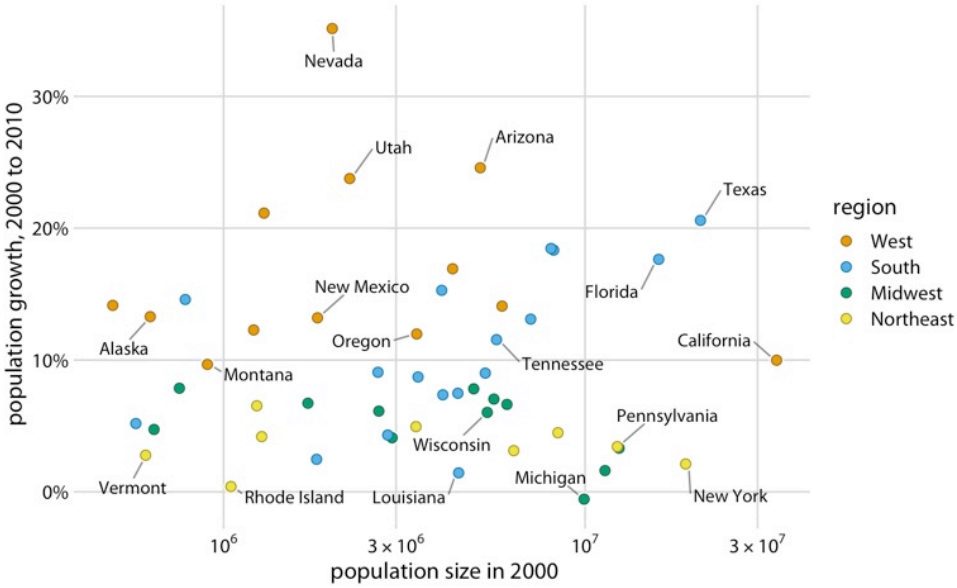
https://clauswilke.com/dataviz/color-pitfalls.html

TOO MANY COLORS

Alternatives:

Direct labelling

Bar chart or other chart type



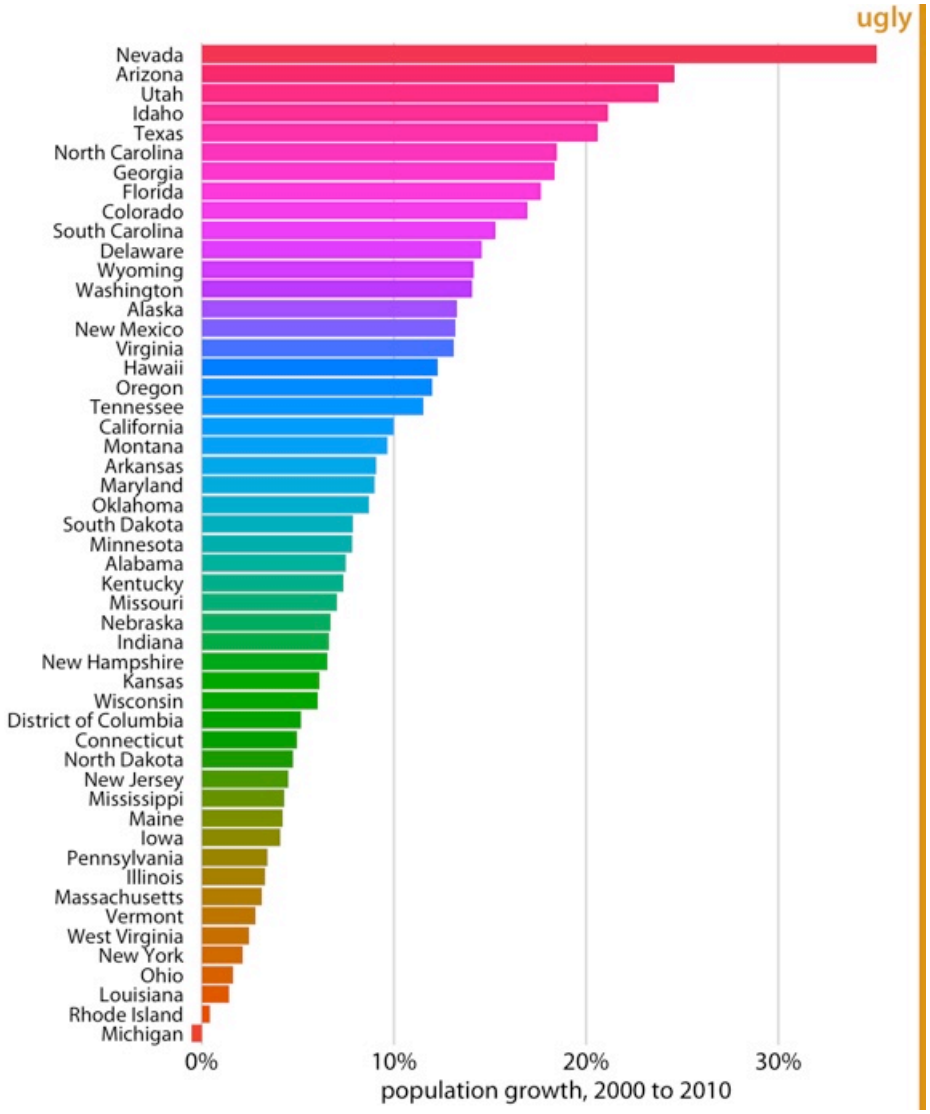
https://clauswilke.com/dataviz/color-pitfalls.html

[10 ways to use fewer colors in your data visualizations]

UNNECESSARY COLOR

When color has no information

Use one color instead

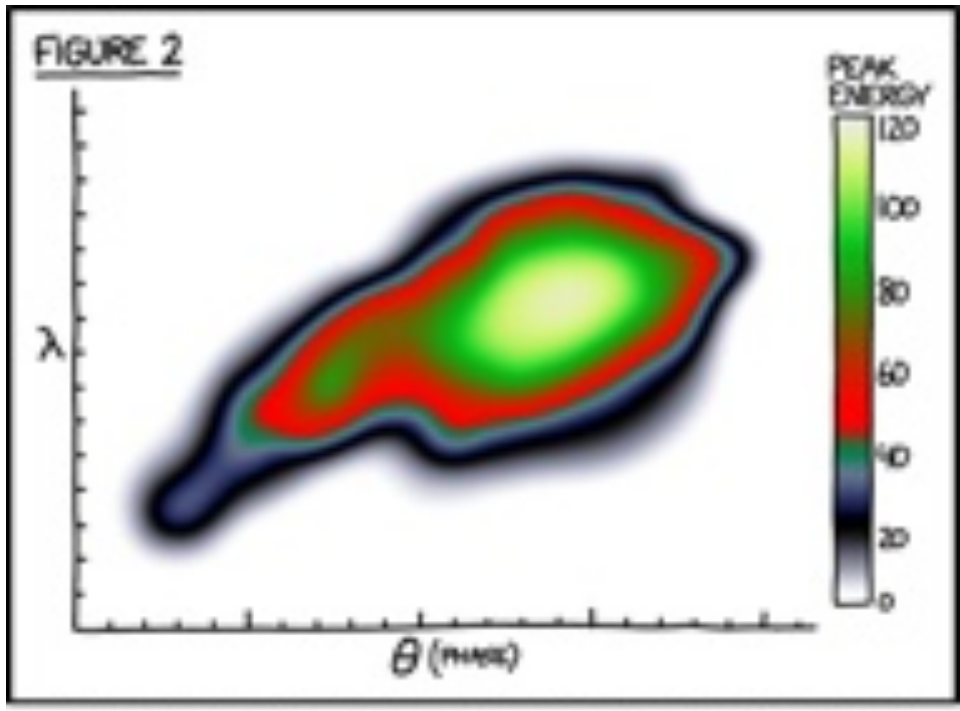


https://clauswilke.com/dataviz/color-pitfalls.html

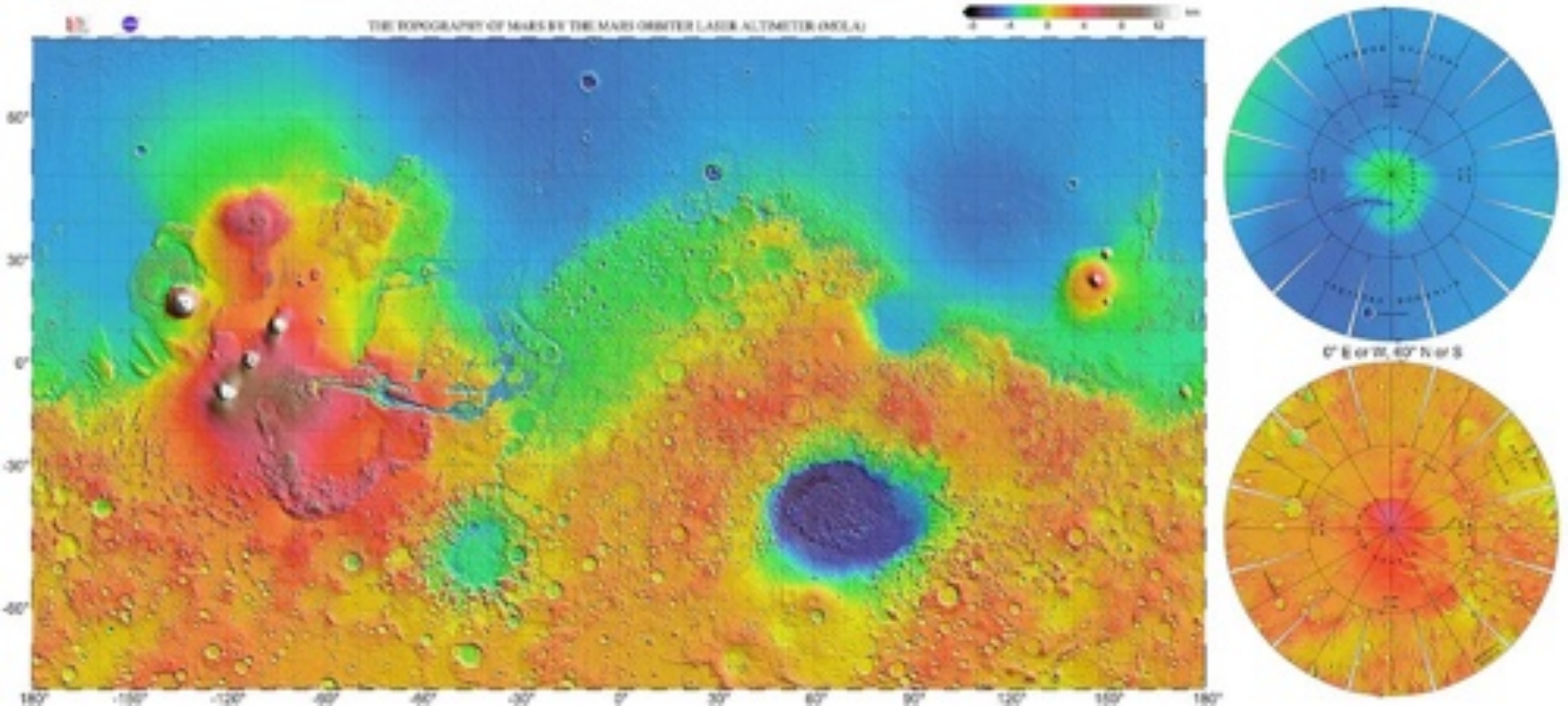
RAINBOW COLOUR SCALES

Caution when using the default rainbow colour scales

How rainbow colour maps can distort data and be misleading



https://xkcd.com/2537/

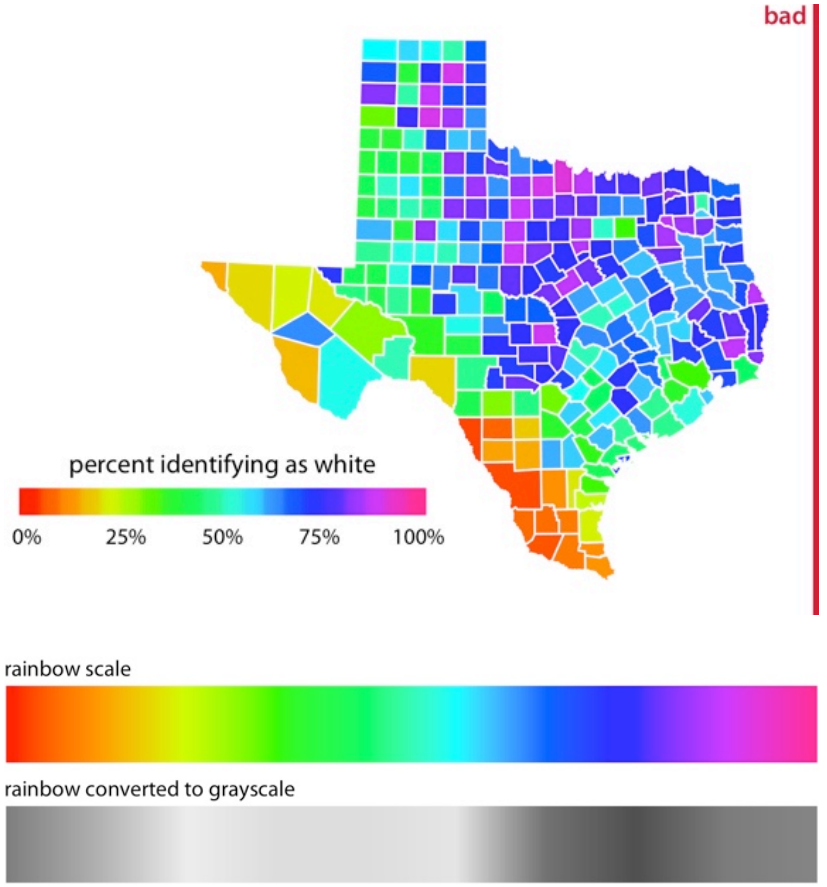


https://attic.gsfc.nasa.gov/mola/images.html

RAINBOW COLOUR MAPS

Difficult to compare values

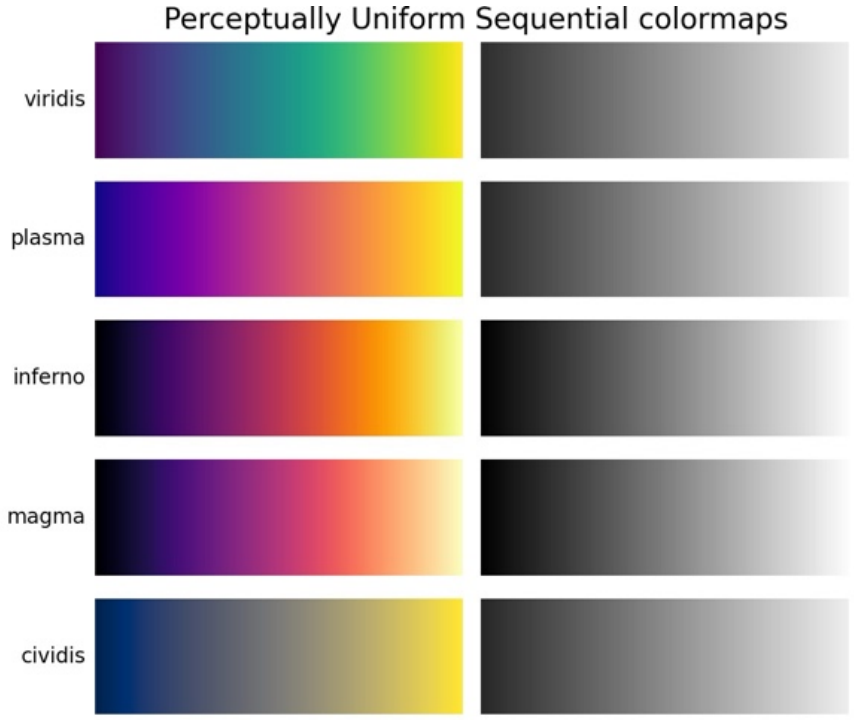
Uneven hues and lightness, misleading



https://clauswilke.com/dataviz/color-pitfalls.html

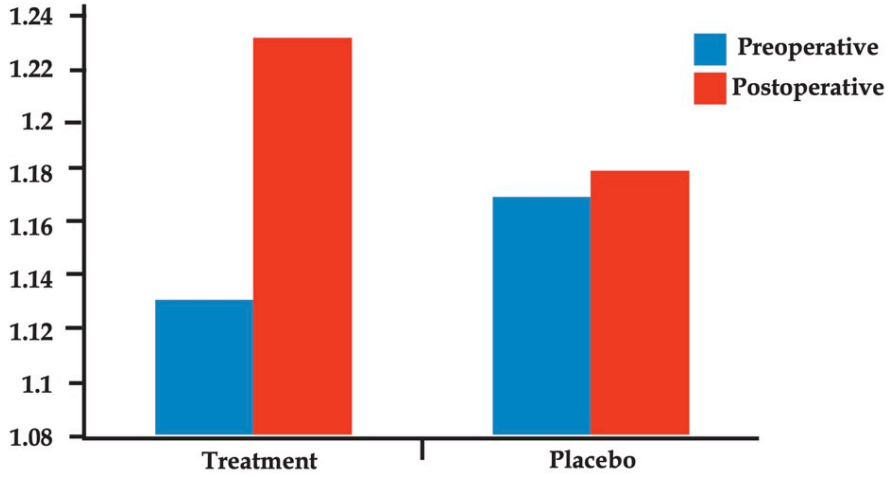
RAINBOW COLOUR SCALES

Use uniform palettes instead



https://matplotlib.org/stable/tutorials/colors/colormaps.html

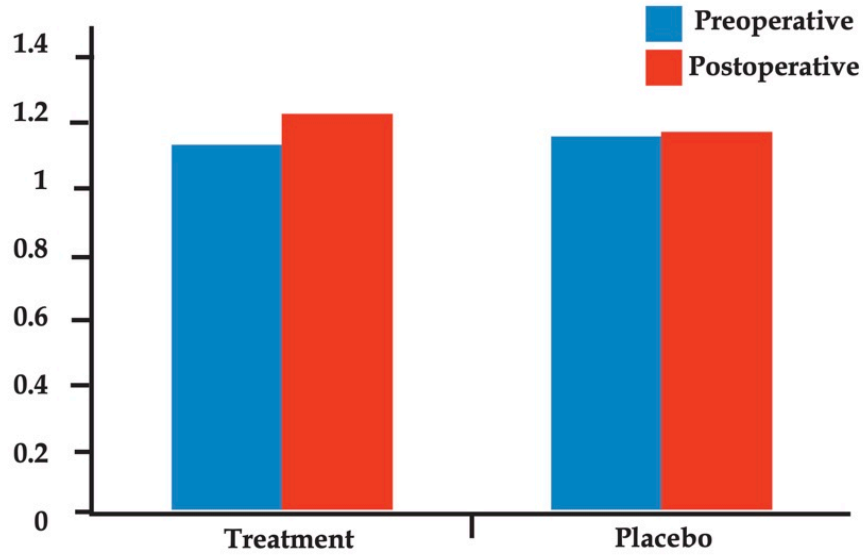
Y-AXIS DOES NOT START AT ZERO



Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014;89(2):280-5

Y-AXIS DOES NOT START AT ZERO

In bar charts y-axis should start at zero



Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014;89(2):280-5

Y-AXIS DOES NOT START AT ZERO

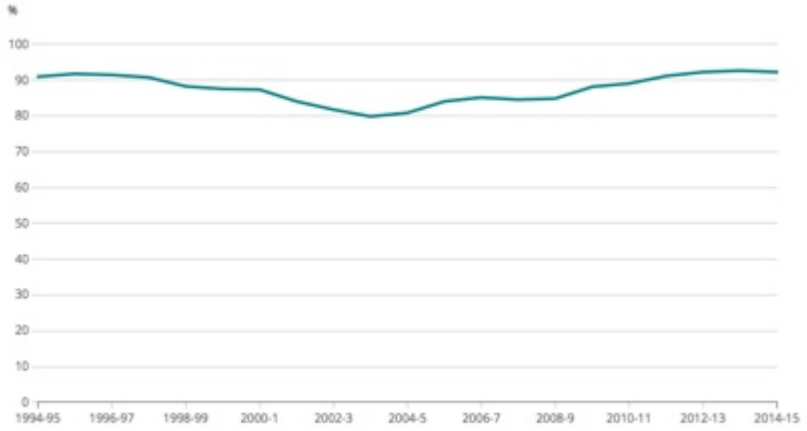
No consensus about other charts

Line chart: same trends but small and important changes could be lost

Where to Start and End Your Y-Axis Scale

COMBINED MMR VACCINATION RATE, 1994-5 TO 2014-15, ENGLAND

Source: NHS Immunisation Statistics - England, 2014-15, Table 8 and 9, HSCIC

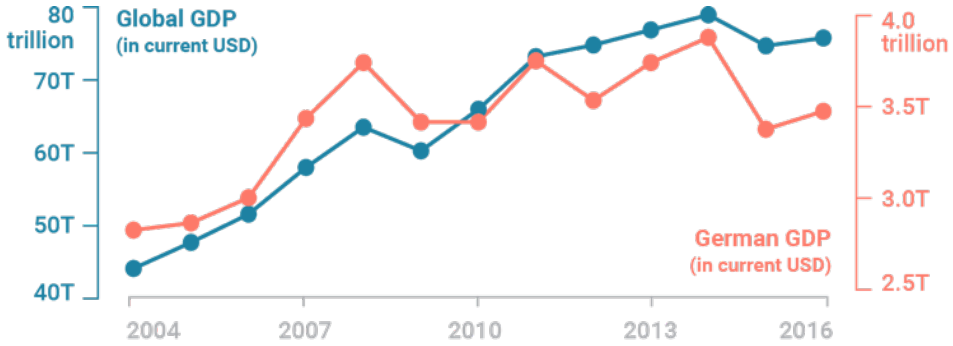


Source: NHS Immunisation Statistics - England, 2014-15, Table 8 and 9, HSCIC

https://digitalblog.ons.gov.uk/2016/06/27/does-the-axis-have-to-start-at-zero-part-1-line-charts/

DUAL Y-AXIS

Avoid

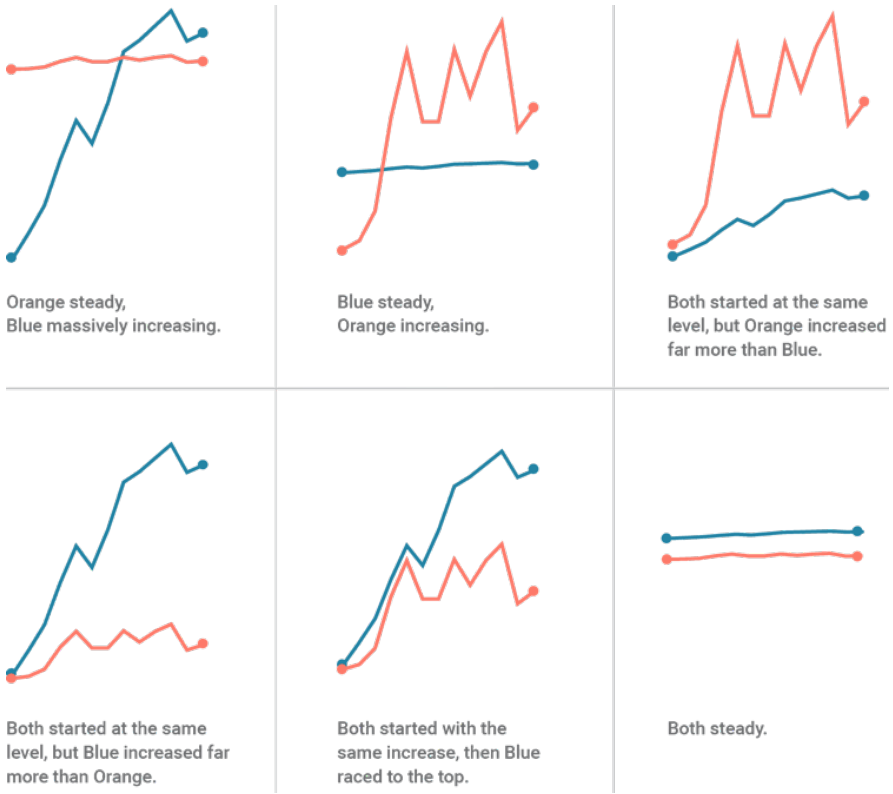


https://blog.datawrapper.de/dualaxis/

DUAL Y-AXIS

Difficult to read

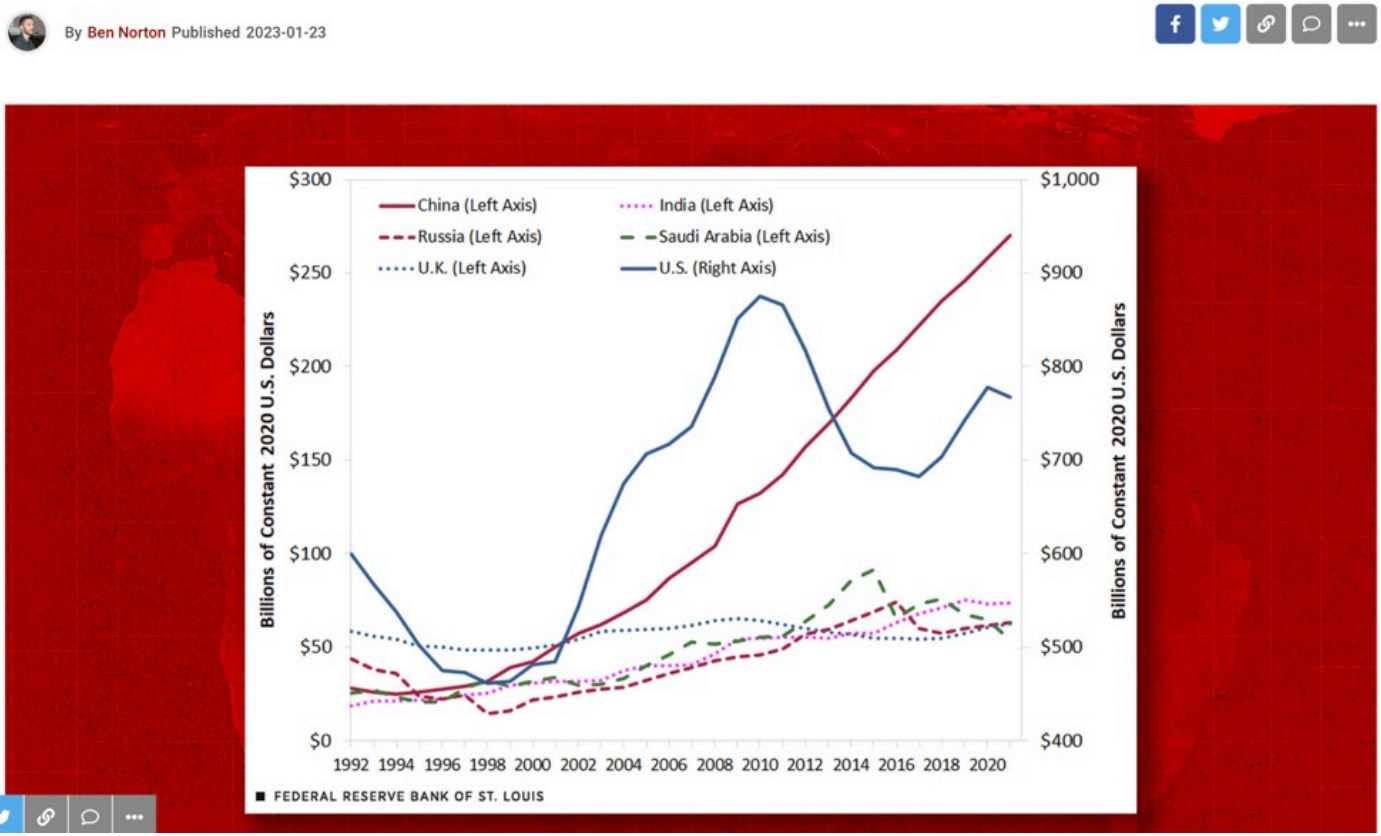
Depending on scaling and position, a correlation or other meaning is implied

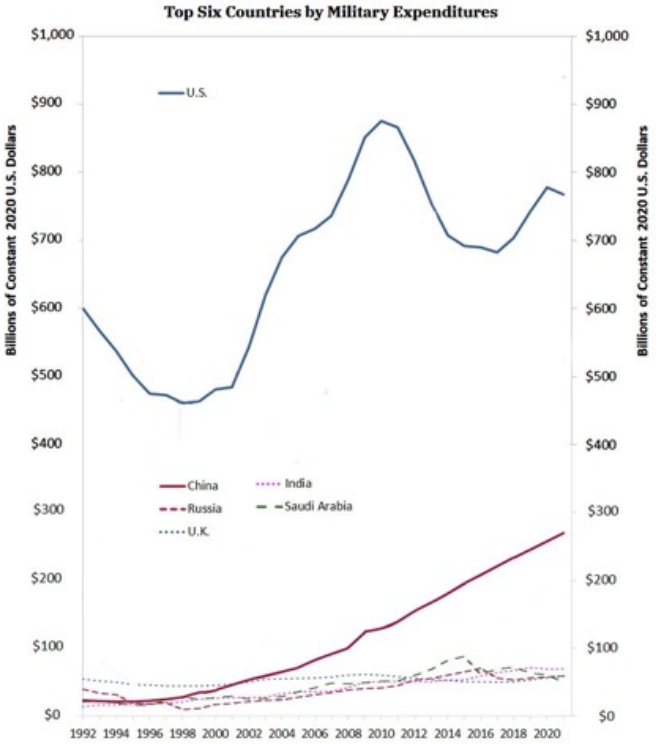


https://blog.datawrapper.de/dualaxis/

Exaggerating China's military spending, St. Louis Fed breaks all statistical rules with misleading graph

The Federal Reserve Bank of St. Louis published a jaw-droppingly misleading graph that portrays China as spending more on its military than the US. In reality, the Pentagon's budget is roughly three times larger.





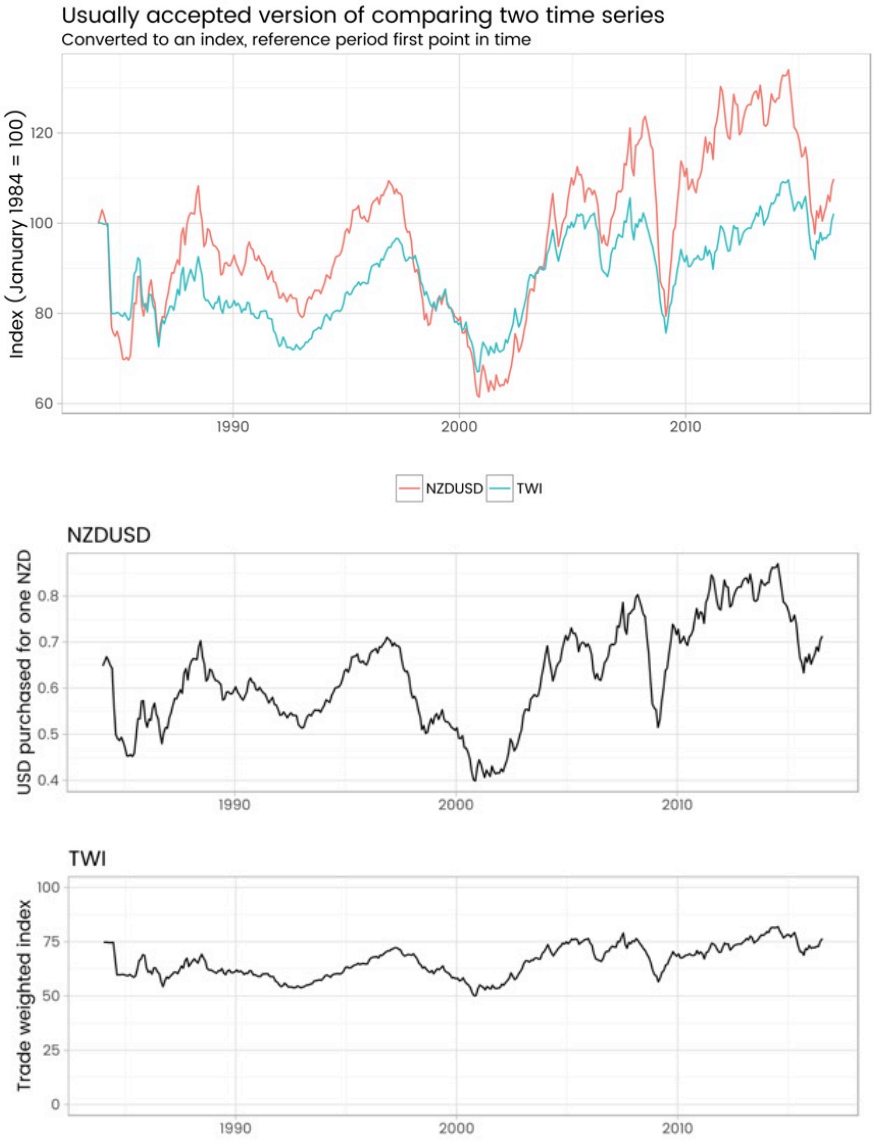
DUAL Y-AXIS

Alternatives:

Indexing

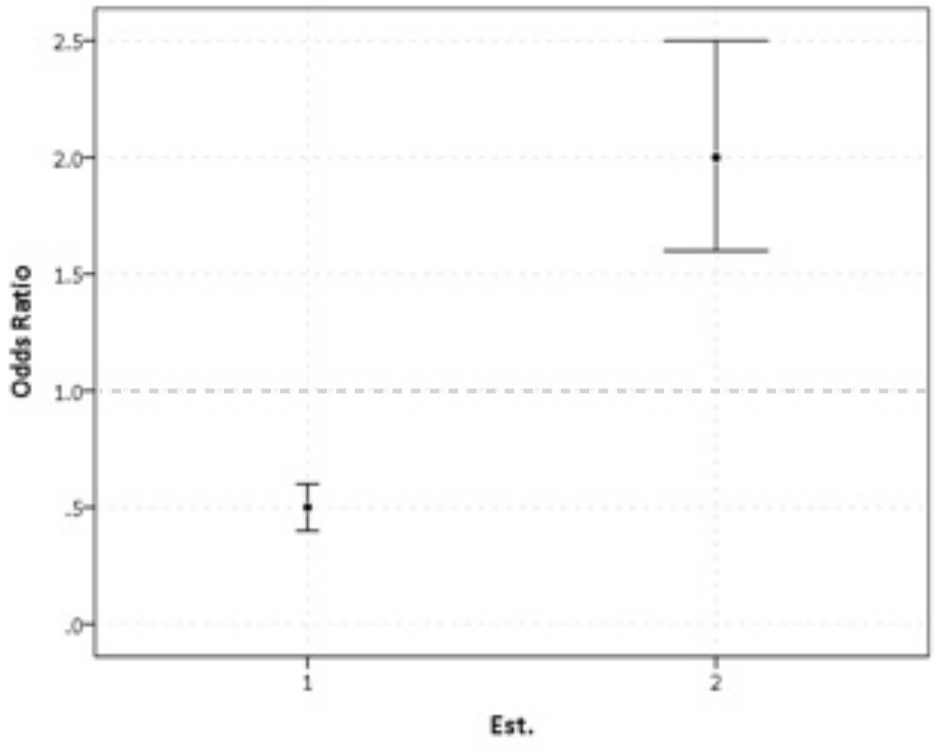
Side by side plots

Why not to use two axes, and what to use instead



http://freerangestats.info/blog/2016/08/18/dualaxes

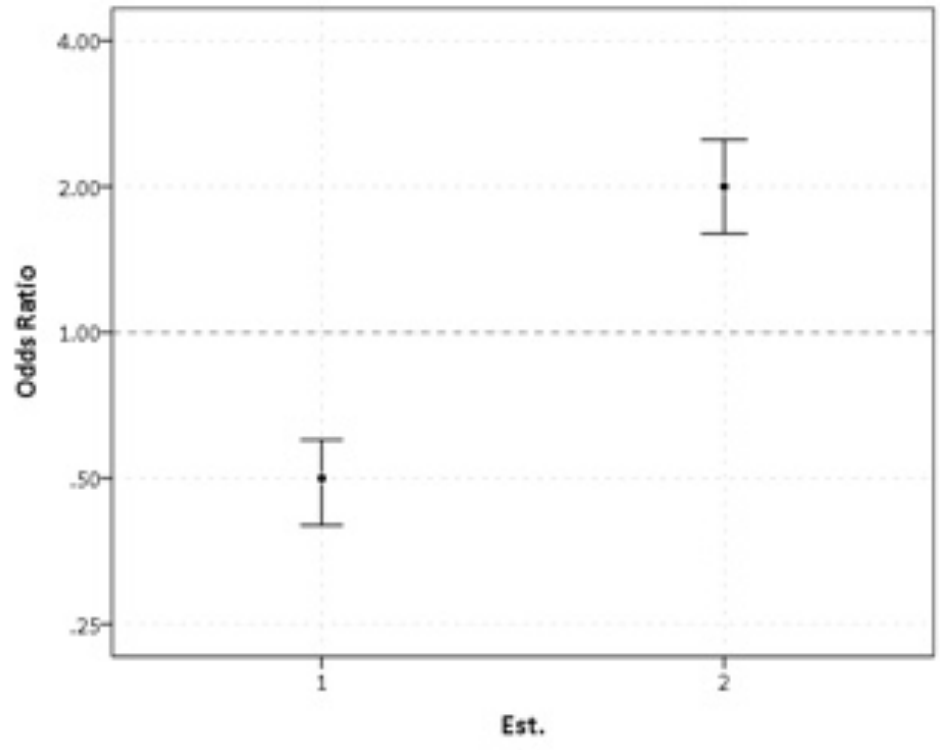
RATIOS THAT DON'T USE LOG AXIS



https://andrewpwheeler.com/2013/10/26/odds-ratios-need-to-be-graphed-on-log-scales/

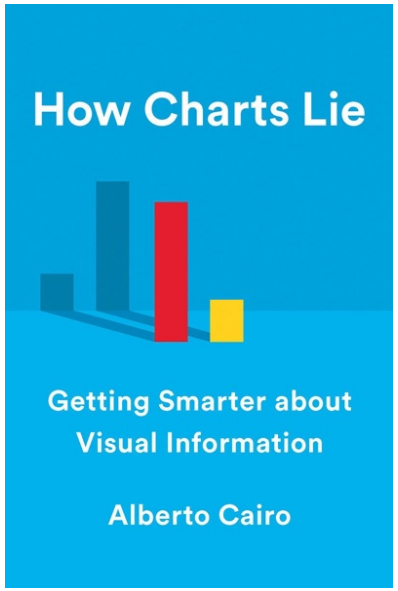
RATIOS THAT DON'T USE LOG AXIS

Use logarithmic scale



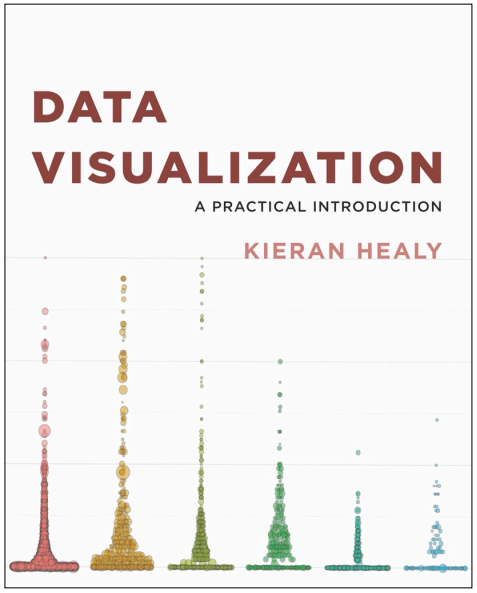
https://andrewpwheeler.com/2013/10/26/odds-ratios-need-to-be-graphed-on-log-scales/

RESOURCES



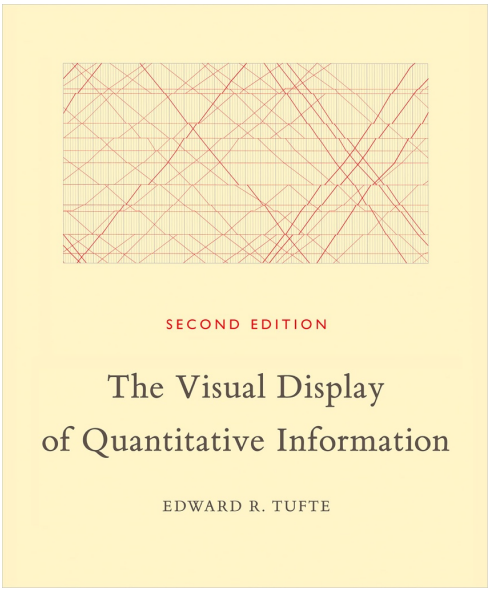
How charts lie: getting smarter about visual information

UU library



Data visualization: a practical introduction

UU library
Free online version



The Visual Display of Quantitative Information

UU library

https://blog.datawrapper.de (theory and how-tos, showcase, stories)

https://nightingaledvs.com (tutorials, stories, industry-related news)

https://flowingdata.com (showcase)

https://r-graph-gallery.com (tutorials with code)

A ggplot2 tutorial for beautiful plotting in R (tutorial with code)

SWD

FIRST, LET'S REVIEW COMMON MYTHS in DATA VIZ

MYTH: LINE GRAPHS are for CONTINUOUS DATA ONLY

The LINES that CONNECT the POINTS HAVE to MAKE SENSE

Example: SURVEY DATA SLOPEGRAPH (a line graph w/ only the points)

MYTH: BARS are ALWAYS BETTER

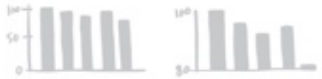
BARS are a GOOD PLACE to START... but NOT ALWAYS the BEST

ASK "what do I want my audience to see?"

Try other types of charts and decide what meets your needs

MYTH: GRAPHS * MUST have a ZERO BASELINE

* THIS is TRUE for BAR CHARTS



MYTH: PIE CHARTS are EVIL

When USING a PIE, ASK yourself WHY?

Studies prove that people read pies & donuts by comparing AREA, not ANGLE

If you think pies make sense for your data & audience, test them out to see!

MYTH: UNBIASED DATA EXISTS

We are BIASING our DATA at EVERY STEP of the PROCESS

RULE! DON'T LIE with DATA

MYTH: MORE DATA is ALWAYS BETTER

BEFORE CHASING after MORE DATA, ask "WHAT will it HELP us DO or DECIDE?"

Audience & context are important when it comes to the right amount of data

MYTH: AVERAGES ALWAYS WORK to SUMMARIZE DATA

YOU NEED to UNDERSTAND the DISTRIBUTION, SPREAD, and VARIABILITY

Averages can mislead by hiding a spread in a single number

MYTH: THERE is a SINGLE RIGHT ANSWER when VISUALIZING DATA

YOU SHOULD ALWAYS CONSIDER when SHOWING DATA: WHAT is your GOAL?

https://www.storytellingwithdata.com/blog/2019/9/30/how-it-came-to-be

DO NOT DECEIVE!

georgios.karamanis@neuro.uu.se

https://karaman.is

geokaramanis

https://github.com/gkaramanis