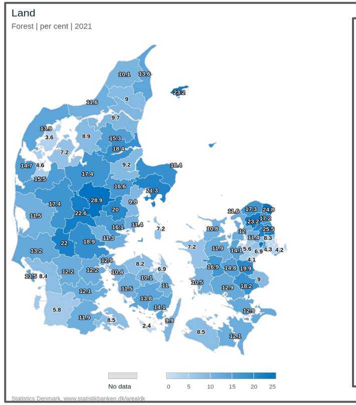
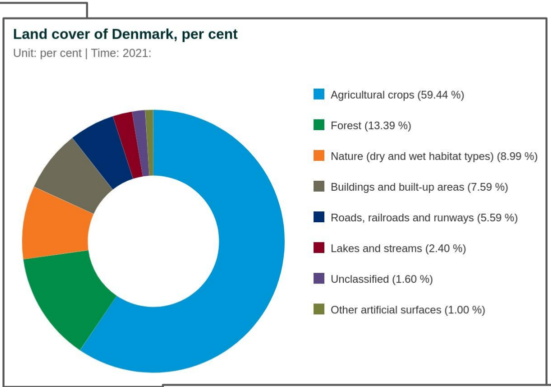
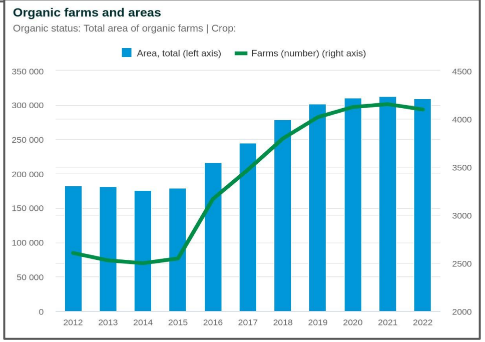
Visualizing Data

Stella Frank stfr@diku.dk

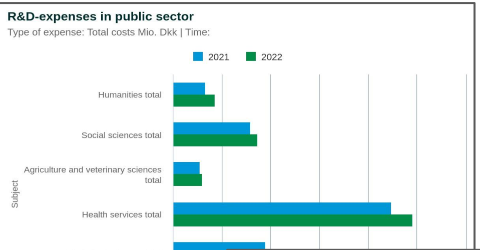
Please have your chart ready, or find one at www.dst.dk

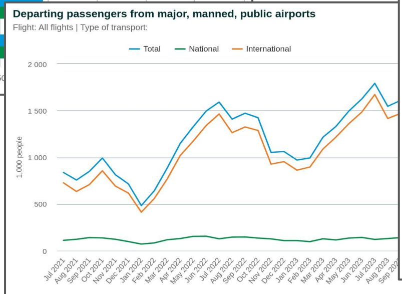


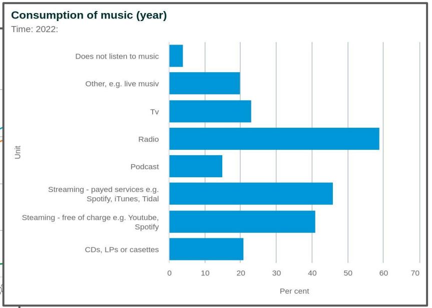




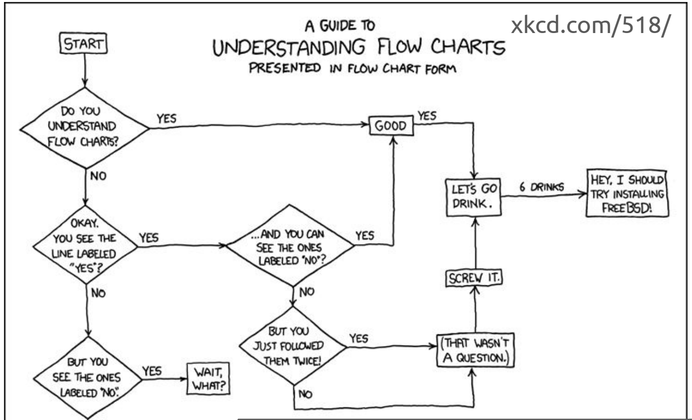
Data Visualizations





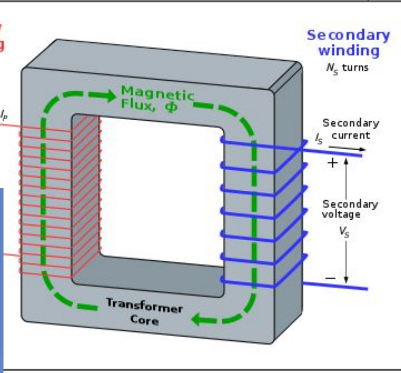


all charts from dst.dk

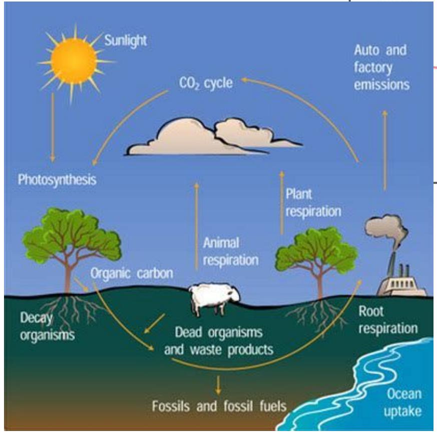

A GUIDE TO
UNDERSTANDING FLOW CHARTS
PRESENTED IN FLOW CHART FORM
START
DO YOU UNDERSTAND FLOW CHARTS?
YES
GOOD
YES
NO
LET'S GO DRINK.
6 DRINKS
HEY, I SHOULD TRY INSTALLING FREEBSD!
OKAY. YOU SEE THE LINE LABELED "YES"?
YES
...AND YOU CAN SEE THE ONES LABELED "NO"?
YES
NO
NO
SCREW IT.
BUT YOU JUST FOLLOWED THEM TWICE!
YES
(THAT WASN'T A QUESTION.)
BUT YOU SEE THE ONES LABELED "NO"?
YES
WAIT, WHAT?
NO
NO
LISTEN.
I HATE YOU.

xkcd.com/518/

<u>Not</u> Data Visualizations



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

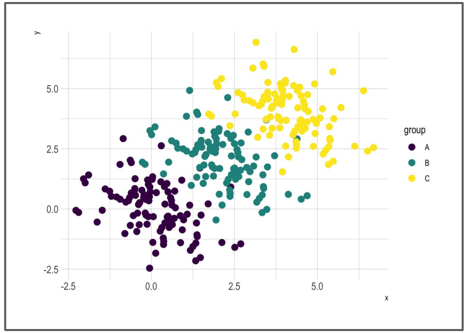


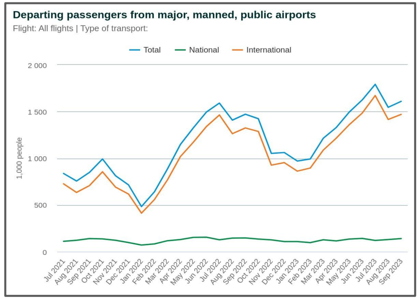
en.wikipedia.org/wiki/Ecosystem_respiration

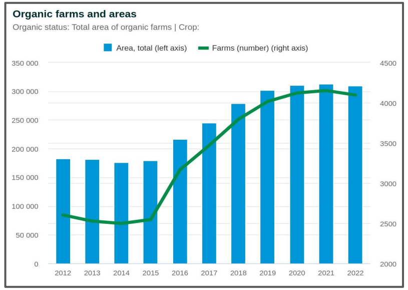
Data visualizations today

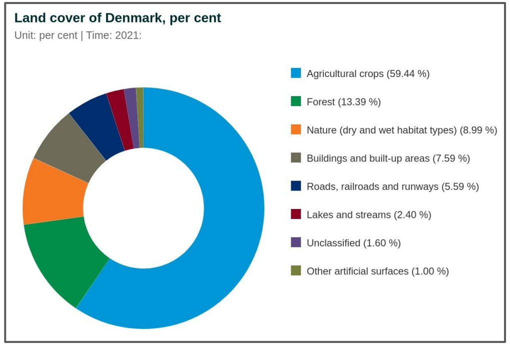
- Frameworks for thinking about different kinds of charts* & when to use them
- Some best practices for visualization design based on human perception
- (Examples of using pandas and seaborn in python.)

*Nomenclature is subtle and inconsistent: chart, graph, plot are ~interchangeable.









more plot

more chart

Visualization is for communication about data

Visualizations go from computer-readable data to human-usable information.

Humans have cognitive limitations:

- Terrible at comparing more than a handful of numbers at once
- Much better at understanding distributions & comparative values graphically
- Still: Bad at comparing more than ~4 variables

Visualizing data can help a lot, but we still need to respect our human limitations.

Example: Numbers vs Graph

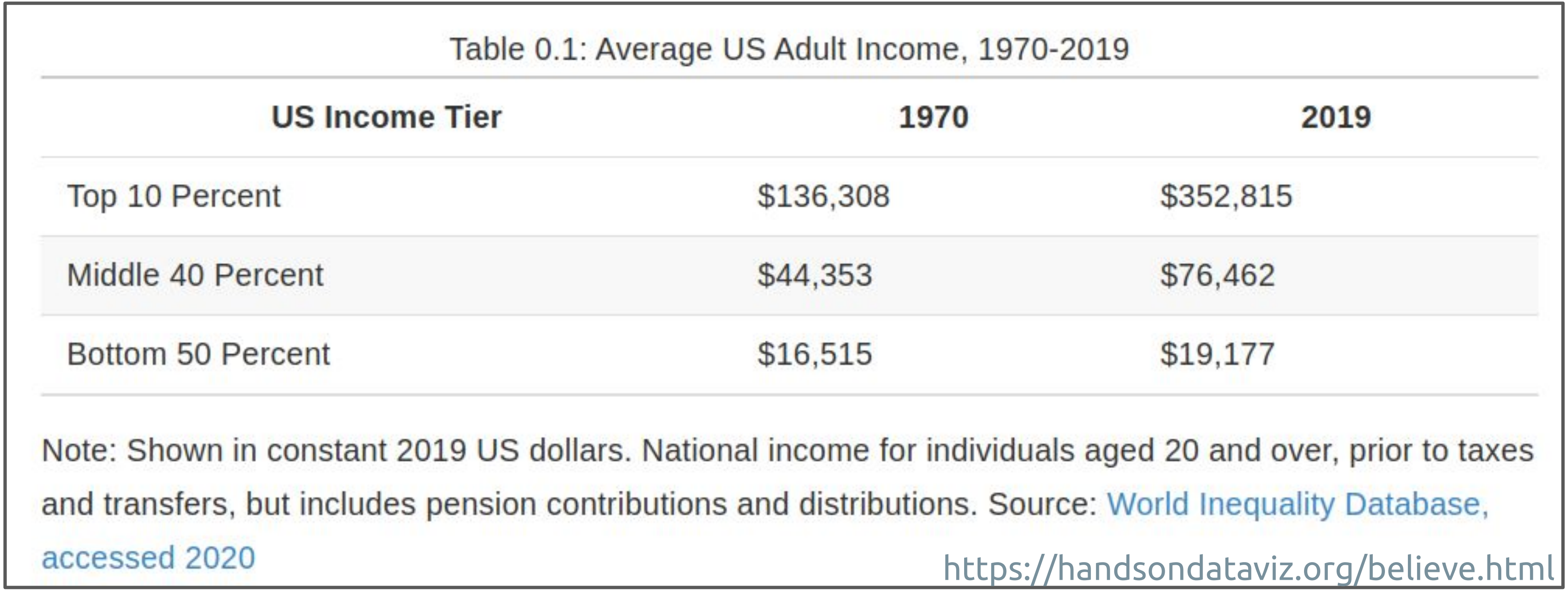


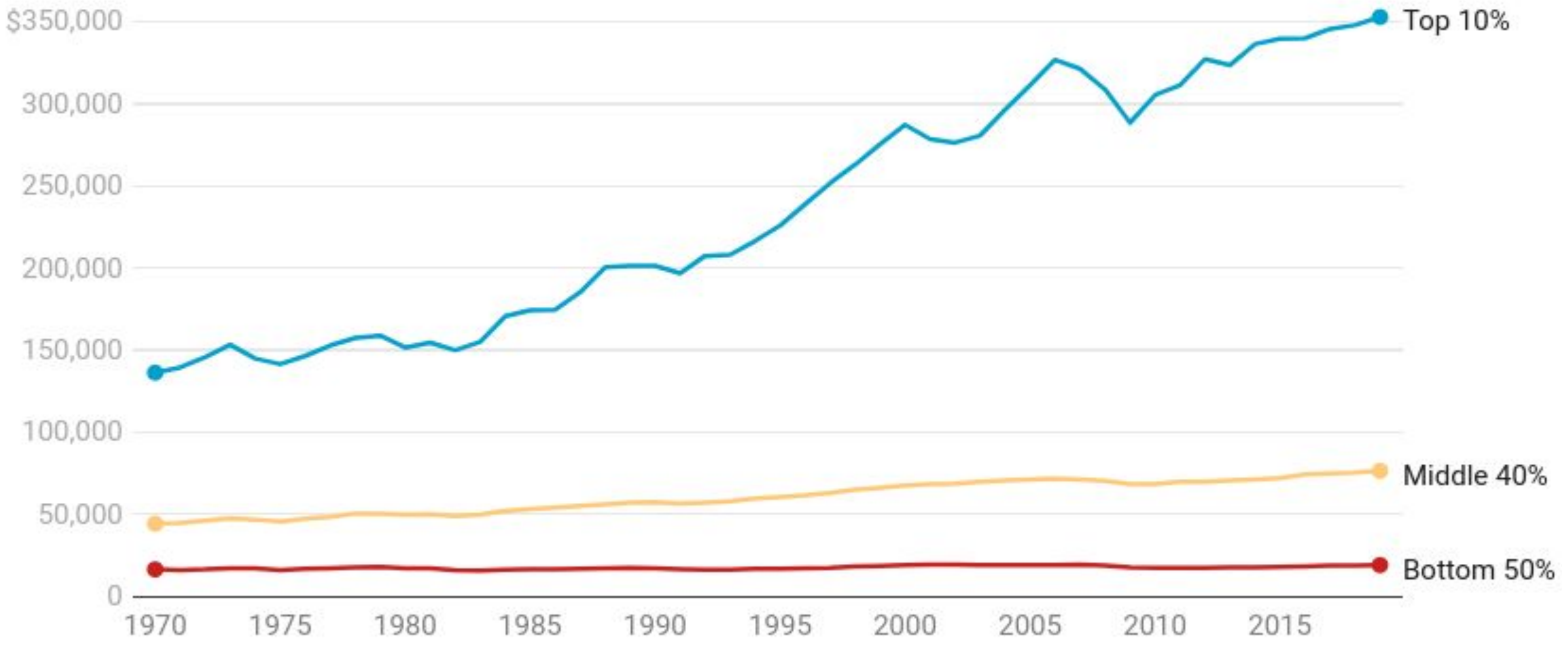
Table 0.1: Average US Adult Income, 1970-2019

US Income Tier	1970	2019
Top 10 Percent	$136,308	$352,815
Middle 40 Percent	$44,353	$76,462
Bottom 50 Percent	$16,515	$19,177

Note: Shown in constant 2019 US dollars. National income for individuals aged 20 and over, prior to taxes and transfers, but includes pension contributions and distributions. Source: World Inequality Database, accessed 2020

https://handsondataviz.org/believe.html

Average US Adult Income, by Percentile, 1970-2019



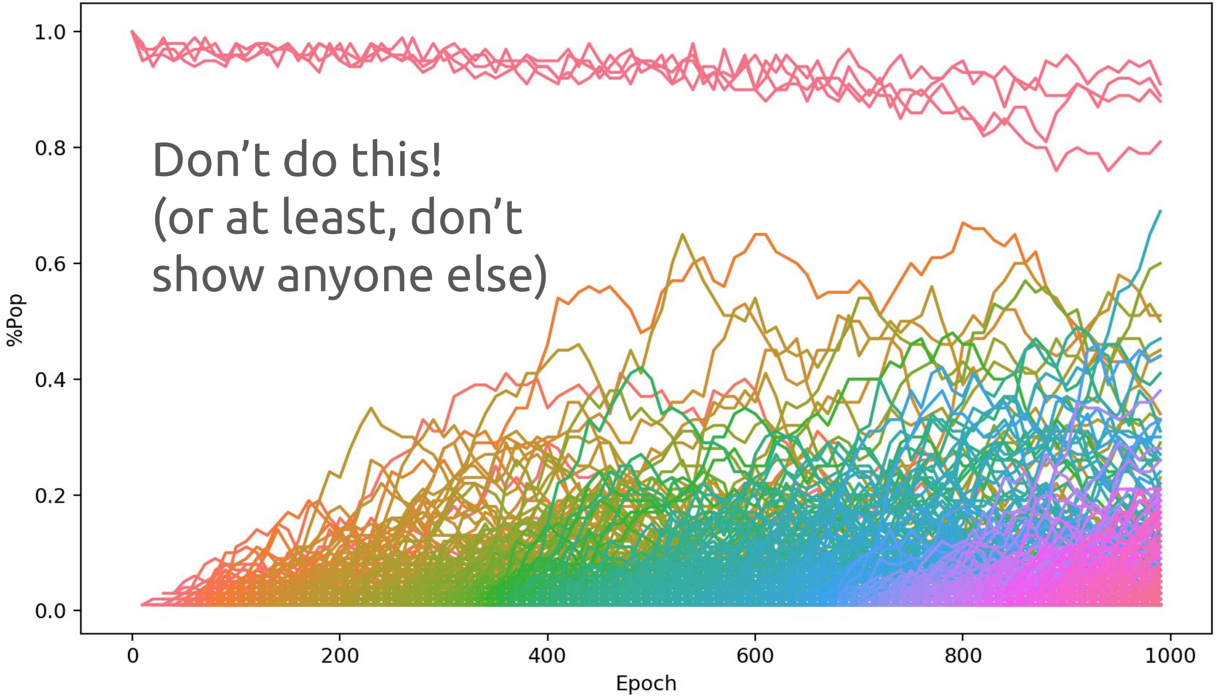
Note: Shown in constant 2019 US dollars. National income for individuals aged 20 and over, prior to taxes and transfers, but includes pension contributions and distributions.

Chart: by HandsOnDataViz • Source: World Inequality Database 2020 • Get the data • Created with Datawrapper

https://handsondataviz.org/believe.html

Good communication requires selection & summarization

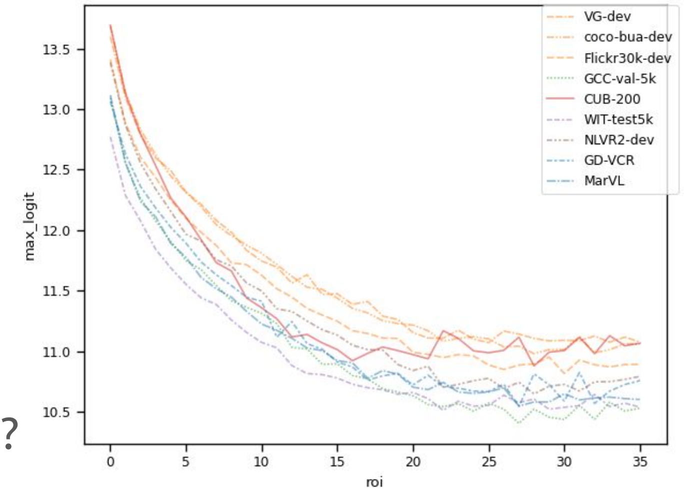
If you plot everything, you show nothing.



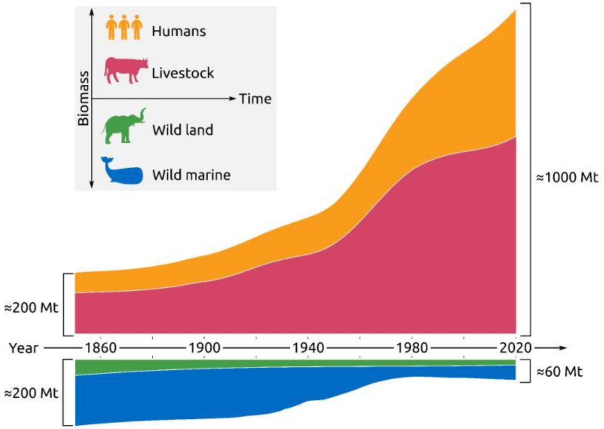
Don’t do this!
(or at least, don’t
show anyone else)

Two types of visualizations

1. **Explorative visualizations**
   Goal: *understand* what is happening in the data.
   - Visualizations as thinking tool; audience = you.
   - Each visualization represents a hypothesis,
   an expectation about what you will see. What is it?



2. **Persuasive visualizations**
   Goal: *convince* your audience that your conclusions are correct.
   - Match between chart/plot & text (data & story)
   - Important to understand & respect your audience's knowledge, attention, defaults.
   Test & revise!



Greenspoon et al, in prep.

Data

What kind of data are represented in your visualizations?

Pairwise exercise, in a minute - but first, let’s think about what data looks like.

Structured vs Unstructured data:

Unstructured: piles of text, images, sounds, DNA

Structured: formatted, classified, ‘structured’ data
most common: tabular data
also: network/graph data, GIS data, etc.

Tabular data - *tidy* data formatting is recommended

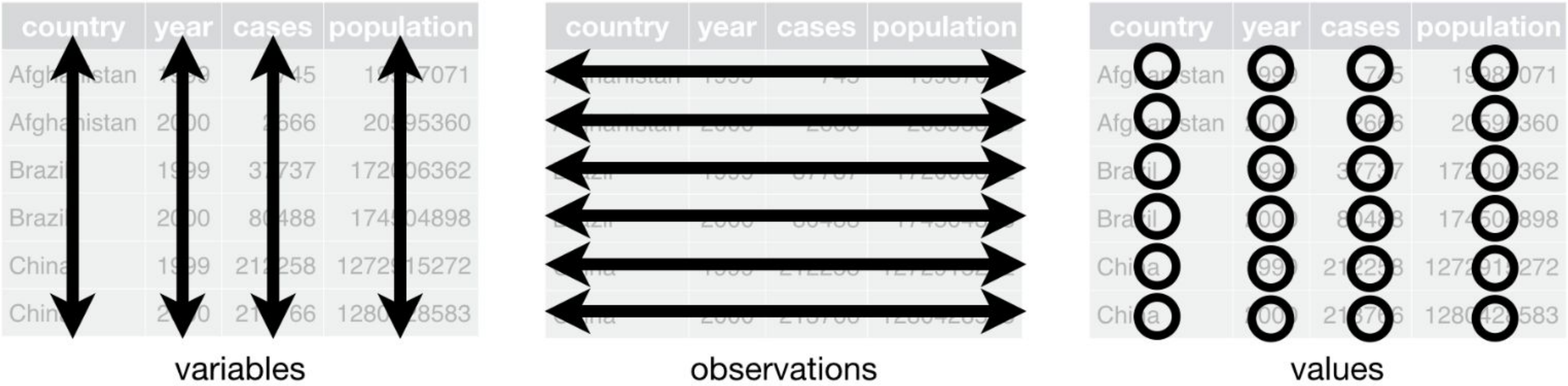


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells.

Wickham & Grolemund, R for Data Science

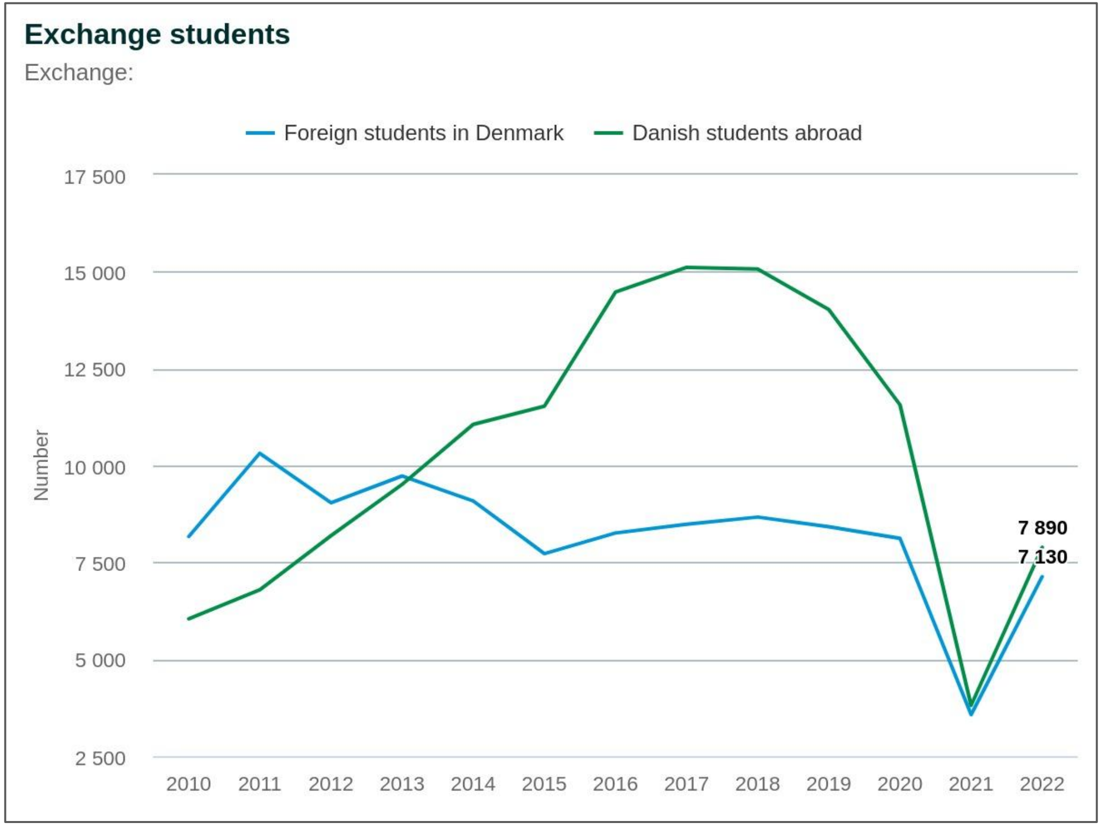
Pairwise, 4 min: What is the data behind your charts?

Look at your charts to determine:

- What are the *observations*?
- What are the *variables?*
- What are the *values?*
- What *units* do the values have?

Does the plot show individual data points, or does it show summarising statistics (e.g. averages)?

Data types: Are the variables' values categorical or quantitative? Discrete or continuous? Maybe even ordinal?



https://www.dst.dk/en/Statistik/emner/uddannelse-og-forskning/fuldtidsuddannelser/udvekslingsstuderende

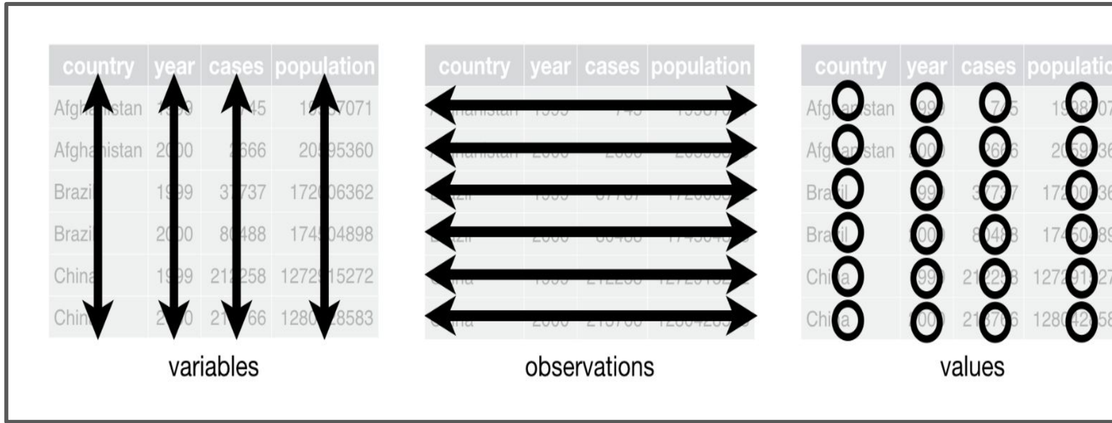
Pairwise, 5 min: What is the data behind your charts?

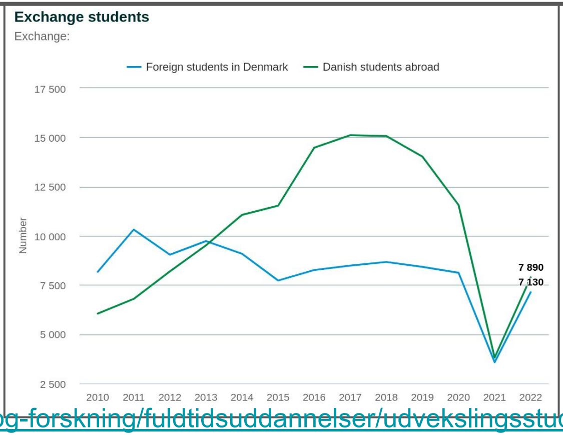
Look at your charts to determine:

- What are the *observations* ?
- What are the *variables?*
- What are the *values?*
- What *units* do the values have?

Does the plot show individual data points, or does it show summarising statistics (e.g. averages)?

Data types: Are the variables' values categorical or quantitative? Discrete or continuous? Maybe even ordinal?




Exchange students

Exchange:

https://www.dst.dk/en/Statistik/emner/uddannelse-og-forskning/fuldtidsuddannelser/udvekslingsstuderende

What did you find?

Have data ⇛ need chart: best practices

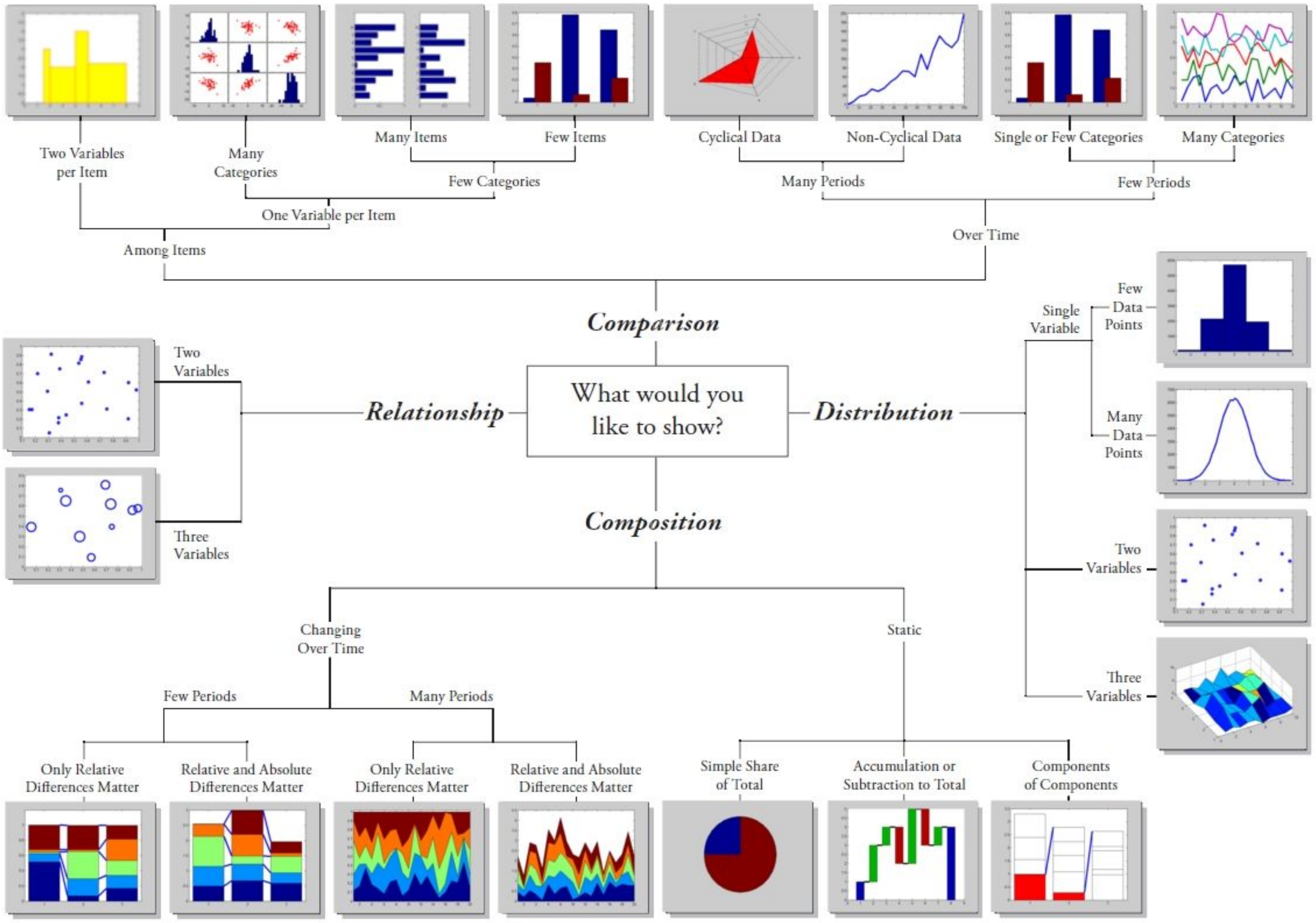
Lots of different kinds of charts - good for different kinds of messages.

Deciding which type of chart to construct depends on data & overall story.

↪ What message is your chart conveying? What is the context?

Two tools to help structure this decision space:
Abela’s Chart Chooser & FT Visual Vocabulary

Abela's Chart Chooser



What would you like to show?

Comparison

Among Items

Two Variables per Item

One Variable per Item

Many Categories

Few Categories

Many Items

Few Items

Over Time

Many Periods

Cyclical Data

Non-Cyclical Data

Few Periods

Single or Few Categories

Many Categories

Relationship

Two Variables

Three Variables

Distribution

Single Variable

Few Data Points

Many Data Points

Two Variables

Three Variables

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Relative and Absolute Differences Matter

Many Periods

Only Relative Differences Matter

Relative and Absolute Differences Matter

Static

Simple Share of Total

Accumulation or Subtraction to Total

Components of Components

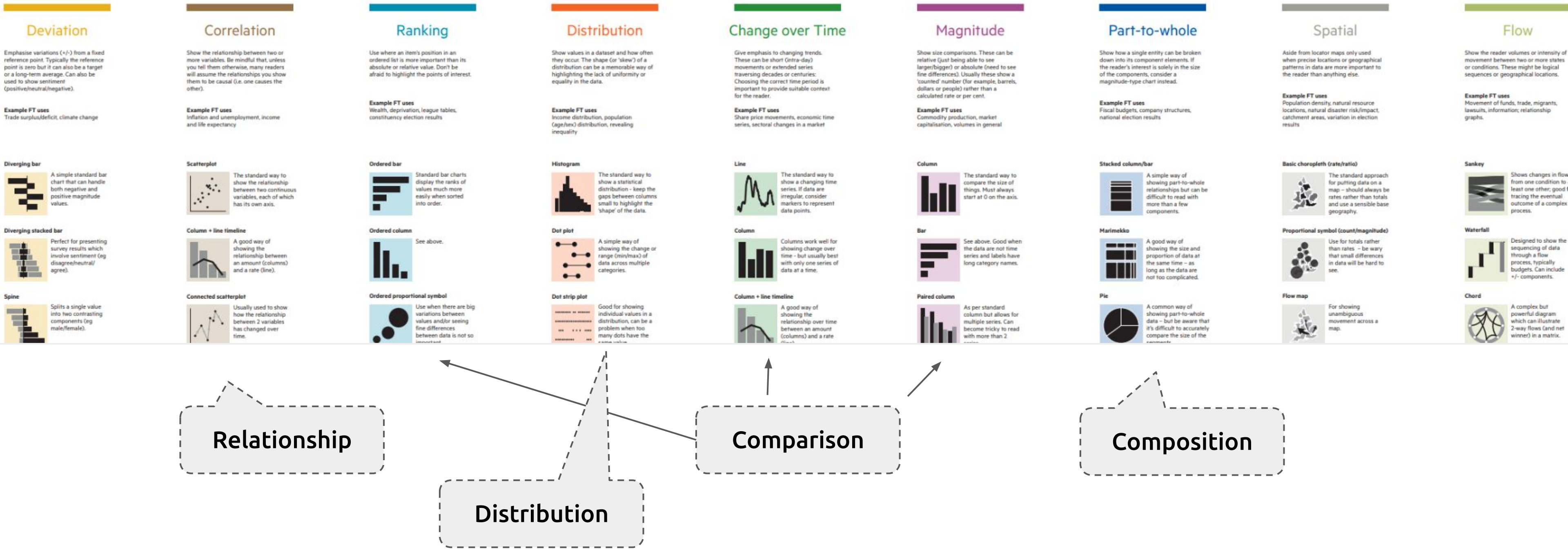
Modified with permision -Doug Hull
blogs.mathworks.com/videos
hull@mathworks.com 2009

www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

FT's Visual Vocabulary

ft-interactive.github.io/visual-vocabulary/

Remix: public.tableau.com/views/VisualVocabulary/VisualVocabulary



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).

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 (columns) and a rate (line).

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

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

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 (age/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

Change over Time

Give emphasis to changing trends. These can be short (intra-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).

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.

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.

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.

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 flow from one condition to least one other; good 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.

Relationship

Comparison

Composition

Distribution

Which category fits your chart?

https://github.com/Financial-Times/chart-doctor/blob/66caa2f126d223b3c68c7360888a035123c4ede6/visual-vocabulary/Visual-vocabulary-en.pdf

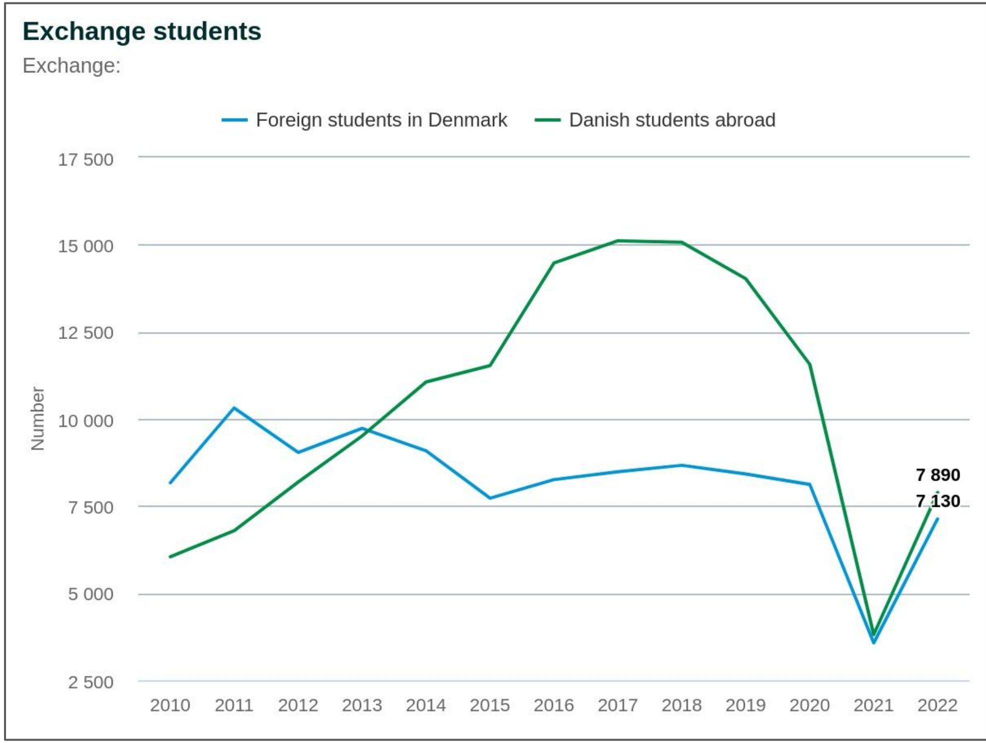
Pairwise, 4 min: What is the story motivating your chart?

Person A: Explain the main point the chart is communicating.
Which VV category does it fall into?

Person B: Do you understand? Are you convinced?

Switch!

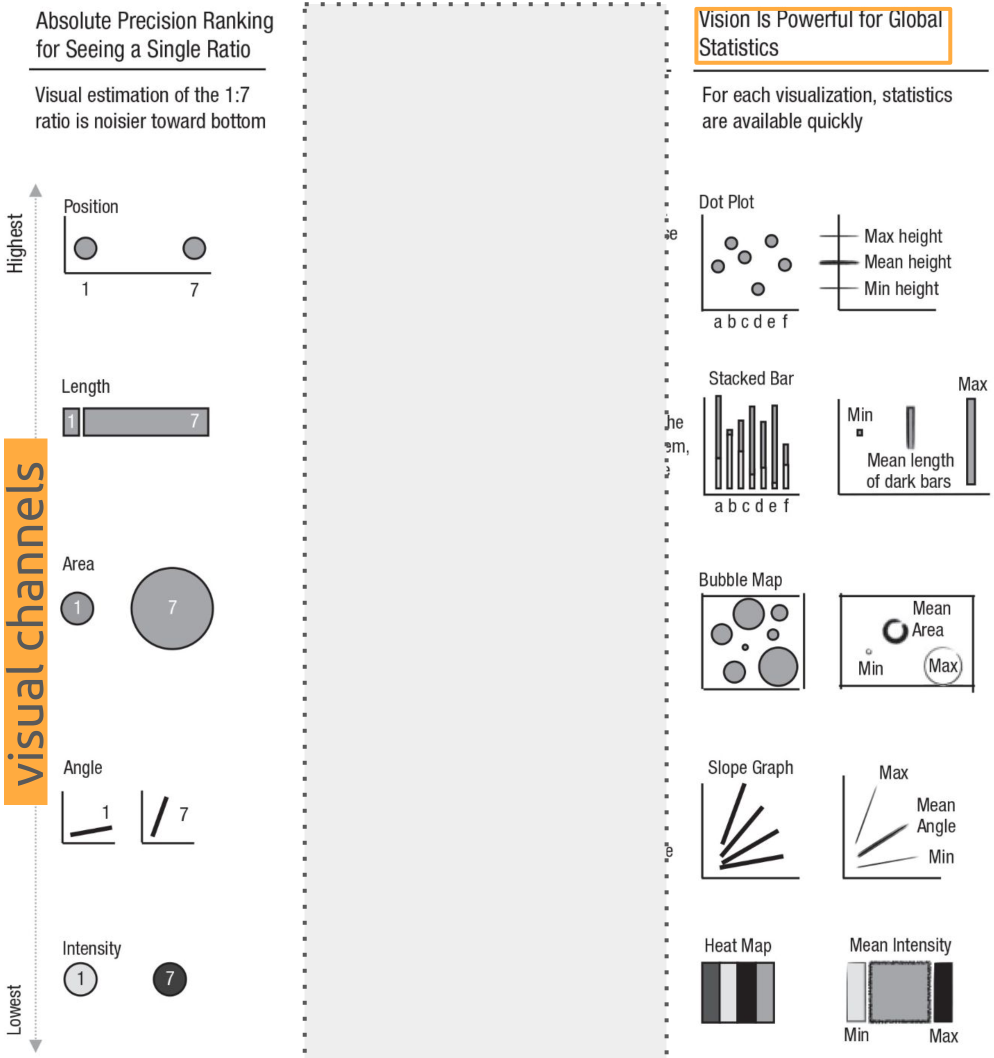
(This might be harder with dst charts: use your imagination.)



https://www.dst.dk/en/Statistik/emner/uddannelse-og-forskning/fuldtidsuddannelser/udvekslingsstuderende

What did you find?

Strengths in Visual Processing



Absolute Precision Ranking for Seeing a Single Ratio

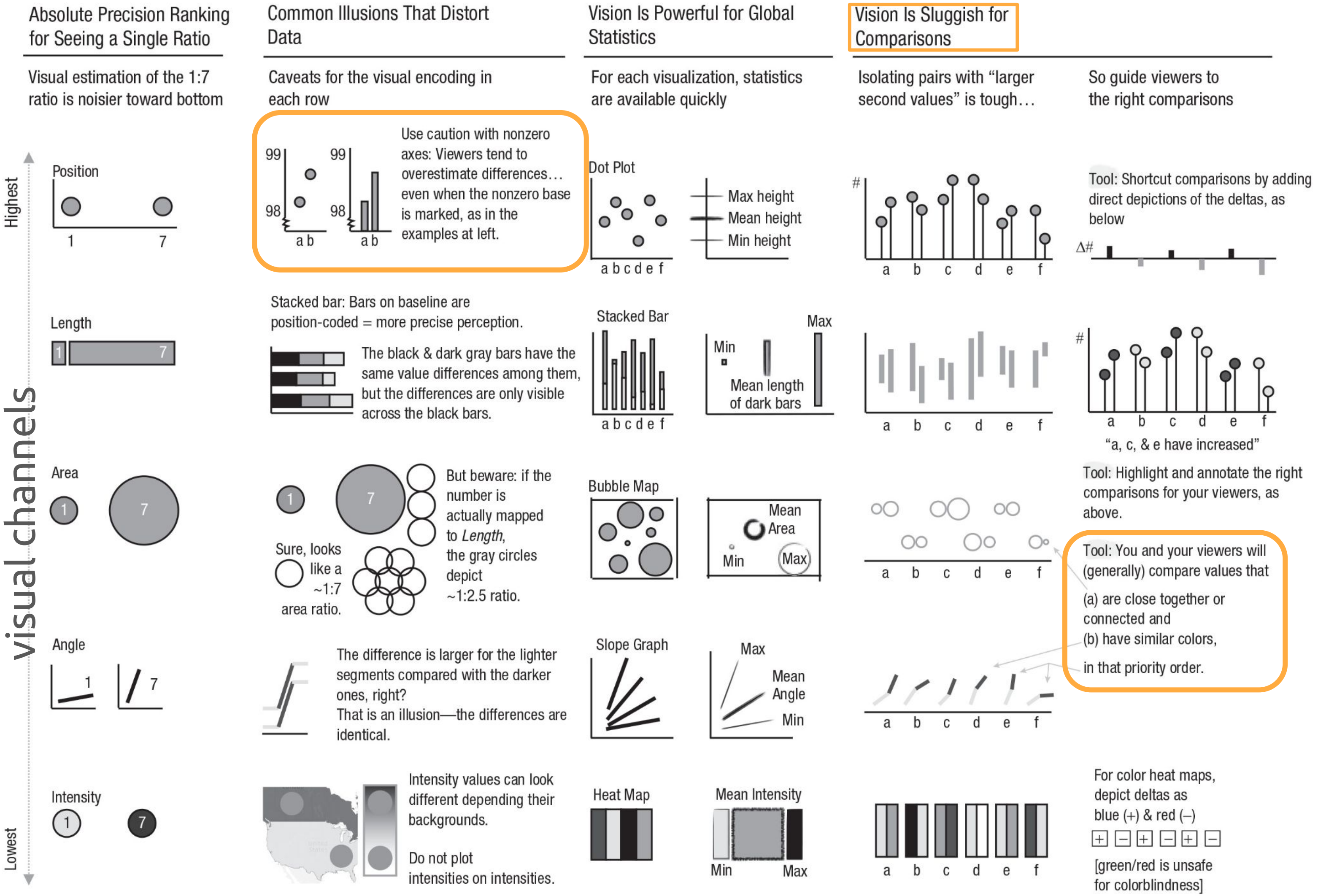
Visual estimation of the 1:7 ratio is noisier toward bottom

Vision Is Powerful for Global Statistics

For each visualization, statistics are available quickly

Figure 2 from:
Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. Psychological Science in the Public Interest, 22(3), 110-161. https://doi.org/10.1177/15291006211051956

Strengths and Weaknesses in Visual Processing



Absolute Precision Ranking for Seeing a Single Ratio

Visual estimation of the 1:7 ratio is noisier toward bottom

Common Illusions That Distort Data

Caveats for the visual encoding in each row

Use caution with nonzero axes: Viewers tend to overestimate differences… even when the nonzero base is marked, as in the examples at left.

Stacked bar: Bars on baseline are position-coded = more precise perception.

The black & dark gray bars have the same value differences among them, but the differences are only visible across the black bars.

But beware: if the number is actually mapped to *Length*, the gray circles depict ~1:2.5 ratio.

The difference is larger for the lighter segments compared with the darker ones, right?
That is an illusion—the differences are identical.

Intensity values can look different depending their backgrounds.

Do not plot intensities on intensities.

Vision Is Powerful for Global Statistics

For each visualization, statistics are available quickly

Vision Is Sluggish for Comparisons

Isolating pairs with "larger second values" is tough…

So guide viewers to the right comparisons

Tool: Shortcut comparisons by adding direct depictions of the deltas, as below

Tool: Highlight and annotate the right comparisons for your viewers, as above.

Tool: You and your viewers will (generally) compare values that
(a) are close together or connected and
(b) have similar colors,
in that priority order.

For color heat maps, depict deltas as blue (+) & red (–)
[green/red is unsafe for colorblindness]

Figure 2 from: Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The Science of Visual Data Communication: What Works. Psychological Science in the Public Interest, 22(3), 110-161. https://doi.org/10.1177/15291006211051956

Bars and Lines

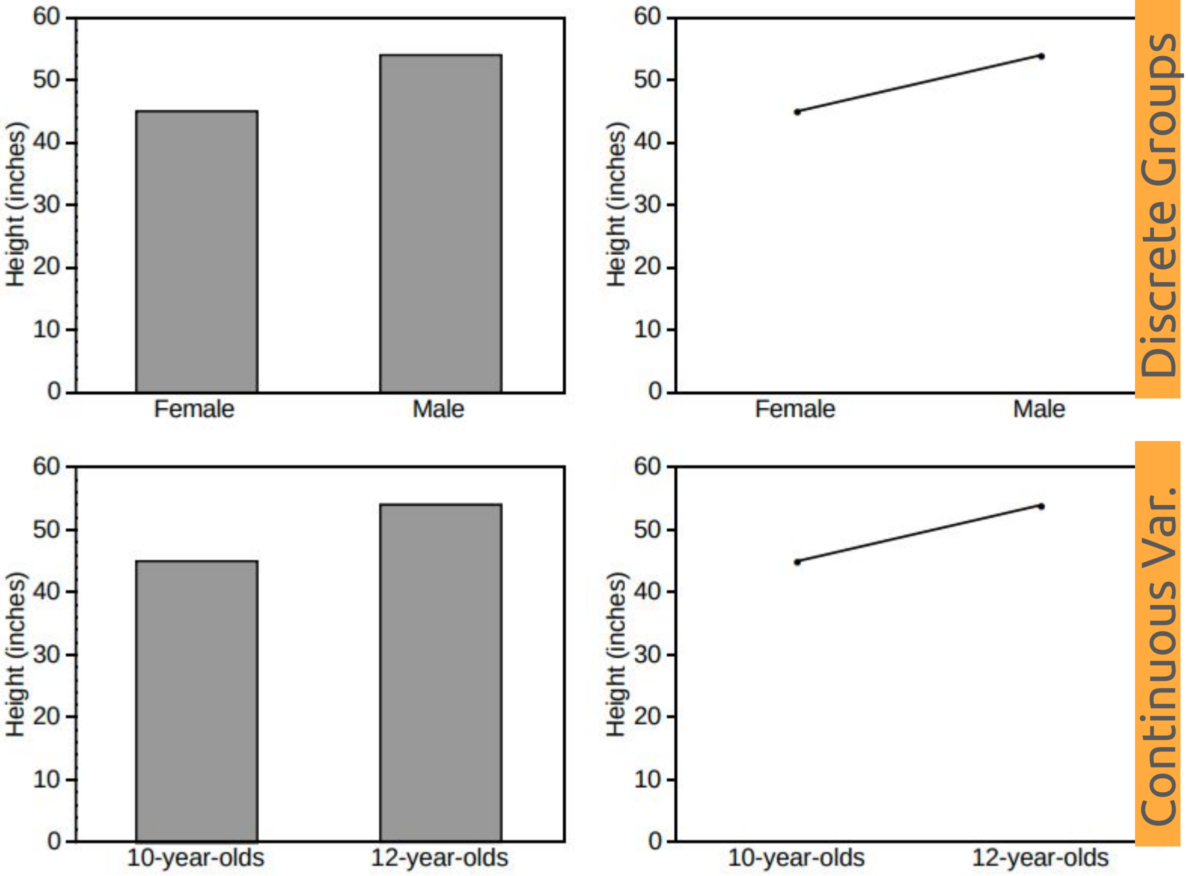
Zacks & Tversky 1999

Experiment: ask people to describe the relationship shown on one of four graphs. Response is ‘discrete comparison’ or ‘trend assessment’?

Results: Mostly congruent BUT: “effect of graph type was about twice that of conceptual domain”

Leading to statements like: “One tends to get taller as one becomes more [male/Danish/only-child]”

Takeaway: use congruent graphs for your data type!



Discrete Groups

Continuous Var.

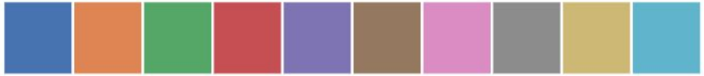
Figure 2. Examples of the bar and line graph stimuli and the continuous and categorical conceptual domains used in Experiment 2.

Discrete areas

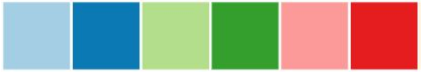
Continuous lines

Color Palettes

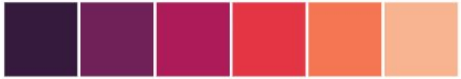
Qualitative - e.g. lines, bars



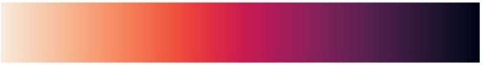
Paired



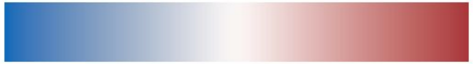
Ordered



Continuous - e.g. heatmaps, maps

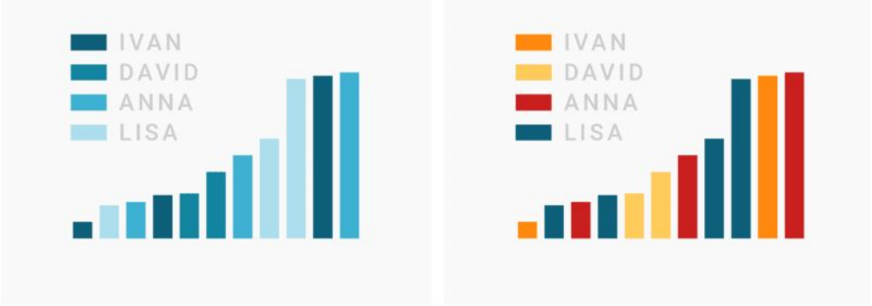


Divergent



https://seaborn.pydata.org/tutorial/color_palettes.html

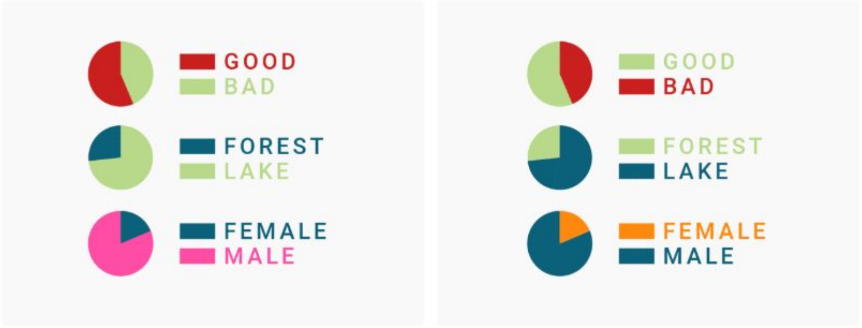
Don’t use gradient palettes for categories (without good reason)



NOT IDEAL

BETTER

Use intuitive colors - match expectations



NOT IDEAL

BETTER

https://blog.datawrapper.de/colors/

Multiple Figures in one Document

When a document has multiple figures, help the reader: **be consistent**

Things that are the **same, stay the same visually:**

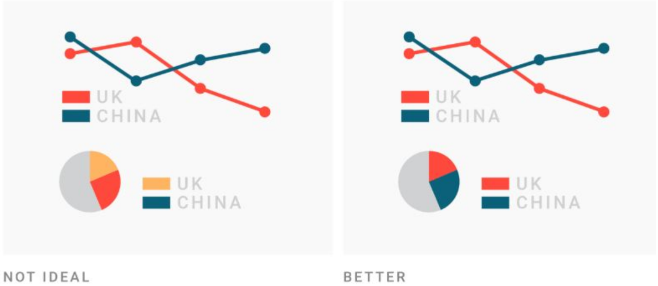
Same observations keep the same colors

e.g. “EU” is always blue;

higher values are always more darker

Same kind of relations use the same kind of chart

e.g. Change over time is always a lineplot



NOT IDEAL

BETTER

Things that are **different look different**

Don’t use the same kind of chart for two different relations

e.g. lineplots for change over time and also change over size

There’s more on Absalon

Python Jupyter notebook using dst data: intro to pandas & seaborn libraries

```
    Files -> L1 Bibliotek notebook
```

Related reading section - see Useful Resources Page on Absalon:

Fundamentals of Data Visualization by Claus O. Wilke

Friends don't let Friends make Bad Graphs by Chenxin Li

& more -

& lots of resources for beginning Python programming.

Next

Thursday lecture: Statistics with Morten Akhøj

Thursday TA sessions: Intro to Python, setup for Assignment 1

Questions?

Admin/logistics/assignment: Daniel Hershkovich - dh@diku.dk

Visualizations/lecture content: Me - stfr@diki.dk