Getting Started with Data Visualisation

Dulcie Vousden, DataKind UK

DISCOVER. LEARN. ANALYSE. SHAPE. REPEAT

Learning objectives

1. Identify factors that make data visualisations less/more successful
2. Understand principles that can be used to design a data visualisation
3. Understand types of visualisations commonly used, and type of data each is best suited to represent
4. Introduce some no-coding tools to get started

Agenda

1. Introductions
2. Exercise: What works/what doesn't
3. 5 steps for data visualisation
4. Thinking about colour
5. Accessibility considerations
6. No-code tools to get started

Exercise: What works, what doesn't?

https://jamboard.google.com/d/1baxUnb4QjNWTtvVB0js4SfqFCxYmX62AV6a58CSBXR0/edit?usp=sharing

Our principles

Steps for an effective data visualization

1. Decide your take home message
2. Define your audience
3. Prep the data
4. Choose the type of visualization
5. Get feedback - Refine!
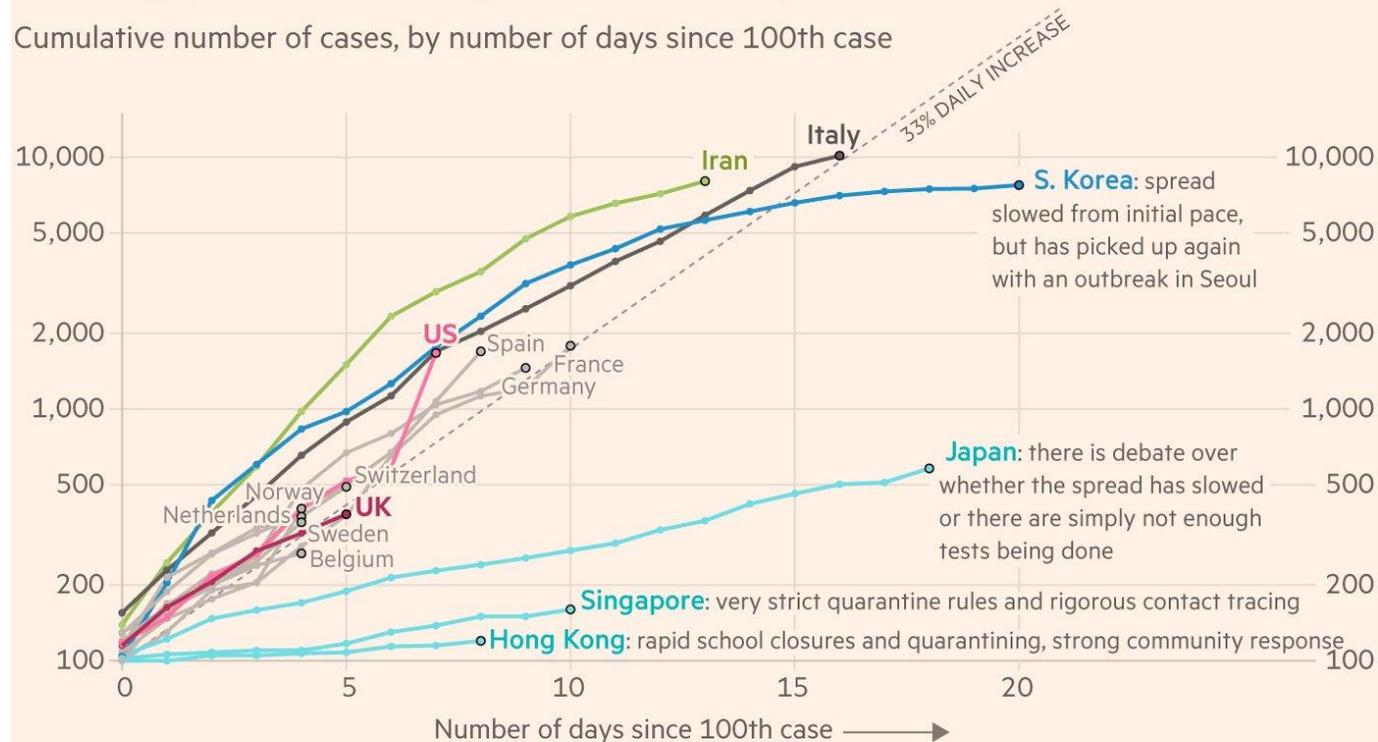
1. Decide your take home message

↘ What would the newspaper headline for this visualisation say?

↘ What do you want your audience to do with this information?

↘ What do you want your audience to conclude?

↘ What action do you want your audience to take?

1. Decide your take home message

Most western countries are on the same coronavirus trajectory. Hong Kong and Singapore have managed to slow the spread

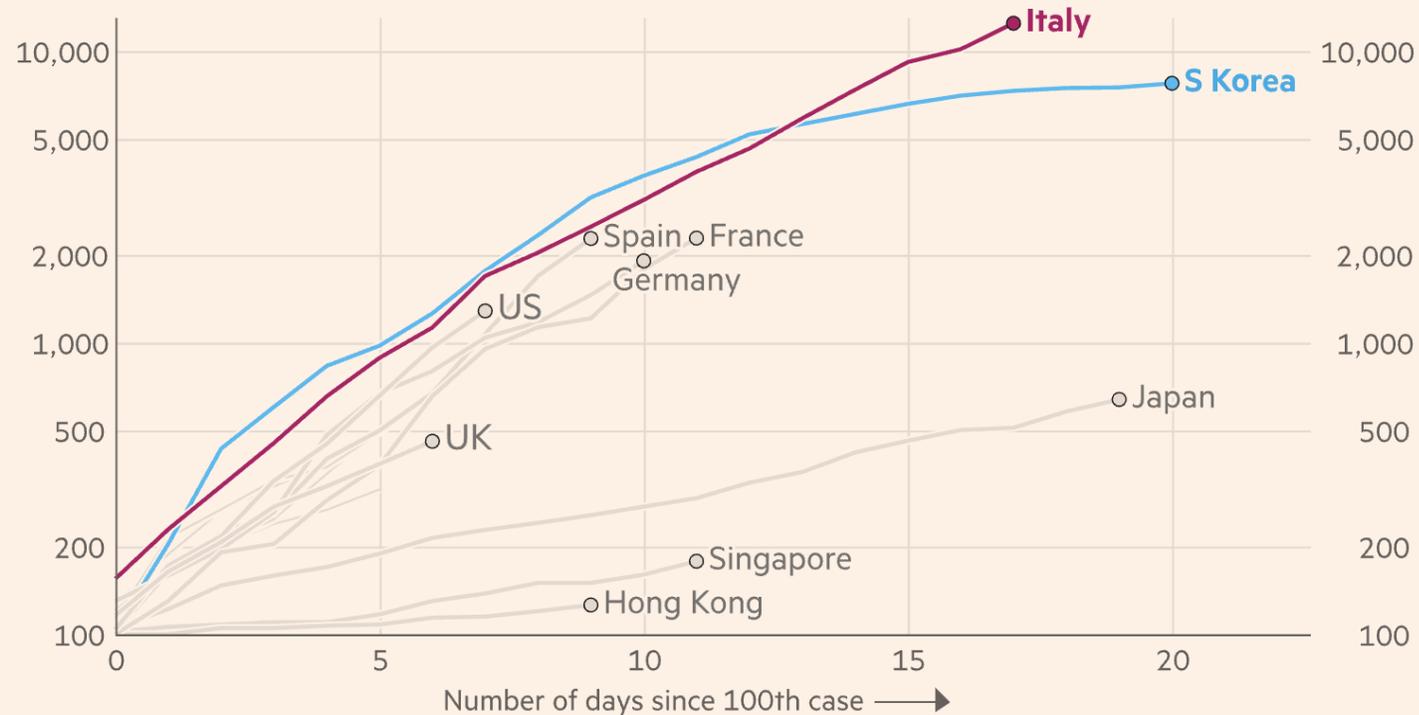Cumulative number of cases, by number of days since 100th case

33% DAILY INCREASE

Italy

Iran

S. Korea: spread slowed from initial pace, but has picked up again with an outbreak in Seoul

US Spain
France
Germany

Switzerland

Norway
Netherlands
UK
Sweden
Belgium

Japan: there is debate over whether the spread has slowed or there are simply not enough tests being done

Singapore: very strict quarantine rules and rigorous contact tracing

Hong Kong: rapid school closures and quarantining, strong community response

10,000
5,000
2,000
1,000
500
200
100

0   5   10   15   20

Number of days since 100th case ⟶

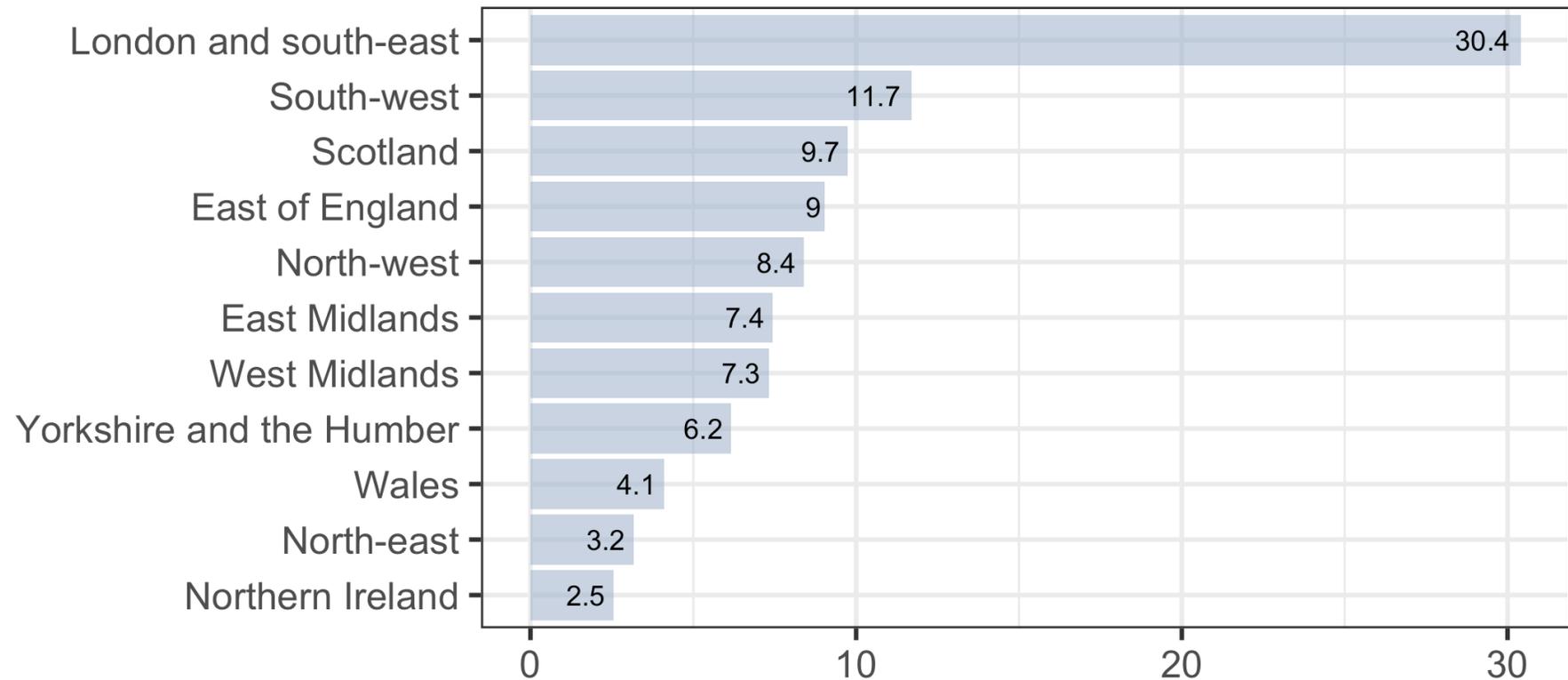https://ig.ft.com/coronavirus-global-data/

1. Decide your take home message

In early March, Italy and South Korea appeared to be on similar paths

Cumulative number of coronavirus cases, by number of days since 100th case (data as of March 12)

Italy
S Korea
Spain France
Germany
US
Japan
UK
Singapore
Hong Kong

Number of days since 100th case ⟶

https://ig.ft.com/coronavirus-global-data/

Voluntary sector employees over-represented in London

Region	Value
London and south-east	30.4
South-west	11.7
Scotland	9.7
East of England	9
North-west	8.4
East Midlands	7.4
West Midlands	7.3
Yorkshire and the Humber	6.2
Wales	4.1
North-east	3.2
Northern Ireland	2.5

Voluntary sector employees by region, June 2019 (%)

https://data.ncvo.org.uk/

2. Define your audience

↘ What background knowledge will this audience have?
↘ Will this be used internally or externally?
↘ What data literacy level will your audience have?
↘ What questions will the audience be looking to answer?
↘ What sorts of decisions will your audience be making with this visualization?

2. Define your audience

Coronavirus in the UK

Total deaths		Total cases	
81,431		**3,072,349**	
Latest daily figure	Two-month trend	Latest daily figure	Two-month trend
563 new deaths		**54,940** new cases	
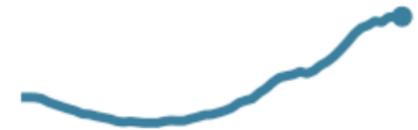
Exercise: Audience & Take-home message

Think-pair-share

- Which audience(s) are you making visualisations for?
- What knowledge will they be bringing?
- What do you want your audience to do with this information?

3. Prep the data

↘ Filter all but the data of interest

↘ Check and clean the data

↘ Perform any summaries needed

↘ What level of detail do you need?

↘ What level of precision do you need?

Data Cleaning checklist

↘ How much data is missing?

↘ Are missing values encoded the same way? (e.g. NA, 'Unknown')

↘ Standardise and validate (e.g. My First Charity > My First Charity Ltd), check postcodes

↘ Are category options consistent (e.g. a computer will treat "Woman" and "woman" as distinct options)

↘ Are dates in the same format (e.g. dd/mm/yyyy vs mm/dd/yyyy)?

↘ Check for plausibility and errors/typos (does an organisation really have an income of £10bn?)

↘ Check for duplicates

↘ Check for PII

4. Choose the type of visualization

↘ What type data do you have?
↘ How much data do you have?
↘ What do you want to show?

What type of data are you using?

Types of data

Numerical data or **quantitative data**
* numerical values

Categorical data or **qualitative data**
* categories

Discrete data
If you can count it, then it is discrete.

Continuous data
If you can measure it, then it is continuous.
* length
* weight
* temperature

Nominal data
If you can brand it, then it is nominal.

Gender
☑ Female
☐ Male

Colour
☐ blue
☑ red
☐ green
☐ orange

Ordinal data
If you can rank or order it, then it is ordinal.
○ Always
◉ Usually
○ Sometimes
○ Rarely
○ Never

How much data do you have?

- How many pieces of data (variables) are you using?
    - **One**: e.g. Client age/gender/ethnicity, amount donated
    - **Two:** e.g. Total fundraising over time,
    - **Three or more:** e.g. Outcome by client sex and age band
- How many observations do you have?
    - **One**: e.g. total donated per individual or total donated per postcode district
    - **Two or more:** e.g. academic performance by school, each year for 10 years; monthly donation per individual

COMPARISON

What would you like to show?

RELATIONSHIP

DISTRIBUTION

COMPOSITION

VARIABLE WIDTH COLUMN CHART — Two Variables per Item

TABLE WITH EMBEDDED CHARTS — Many Categories

BAR CHART HORIZONTAL

BAR CHART VERTICAL — Few Categories

One Variable per Item

Among Items

CIRCULAR AREA CHART — Cyclical Data

LINE CHART — Non-Cyclical Data

BAR CHART VERTICAL — Single or Few Categories

LINE CHART — Many Categories

Over Time

COMPARISON

SCATTER PLOT — Two Variables

SCATTER PLOT BUBBLE SAZE — Three or more Variables

RELATIONSHIP

What would you like to show?

DISTRIBUTION

BAR HISTOGRAM — Few Data Points

Single Variable

LINE HISTOGRAM

Many Data Points

SCATTER PLOT

Two Variables

COMPOSITION

Changing Over Time — Few Periods, Many Periods

Static

Few Periods:
- Only Relative Differences Matter → STACKED 100% BAR CHART
- Relative and Absolute Differences Matter → STACKED BAR CHART

Many Periods:
- Only Relative Differences Matter → STACKED AREA 100% CHART
- Relative and Absolute Differences Matter → STACKED AREA CHART

Static:
- Simple Share of Total → PIE CHART
- Accumulation or Subtraction to Total → WATERFALL CHART
- Components of Components → STACKED 100% BAR CHART WITH SUBCOMPONENTS
- Accumulation to total and absolute difference matters → TREE MAP

CHART SUGGESTIONS - A THOUGHT-STARTER

https://infogram.com/page/choose-the-right-chart-data-visualization
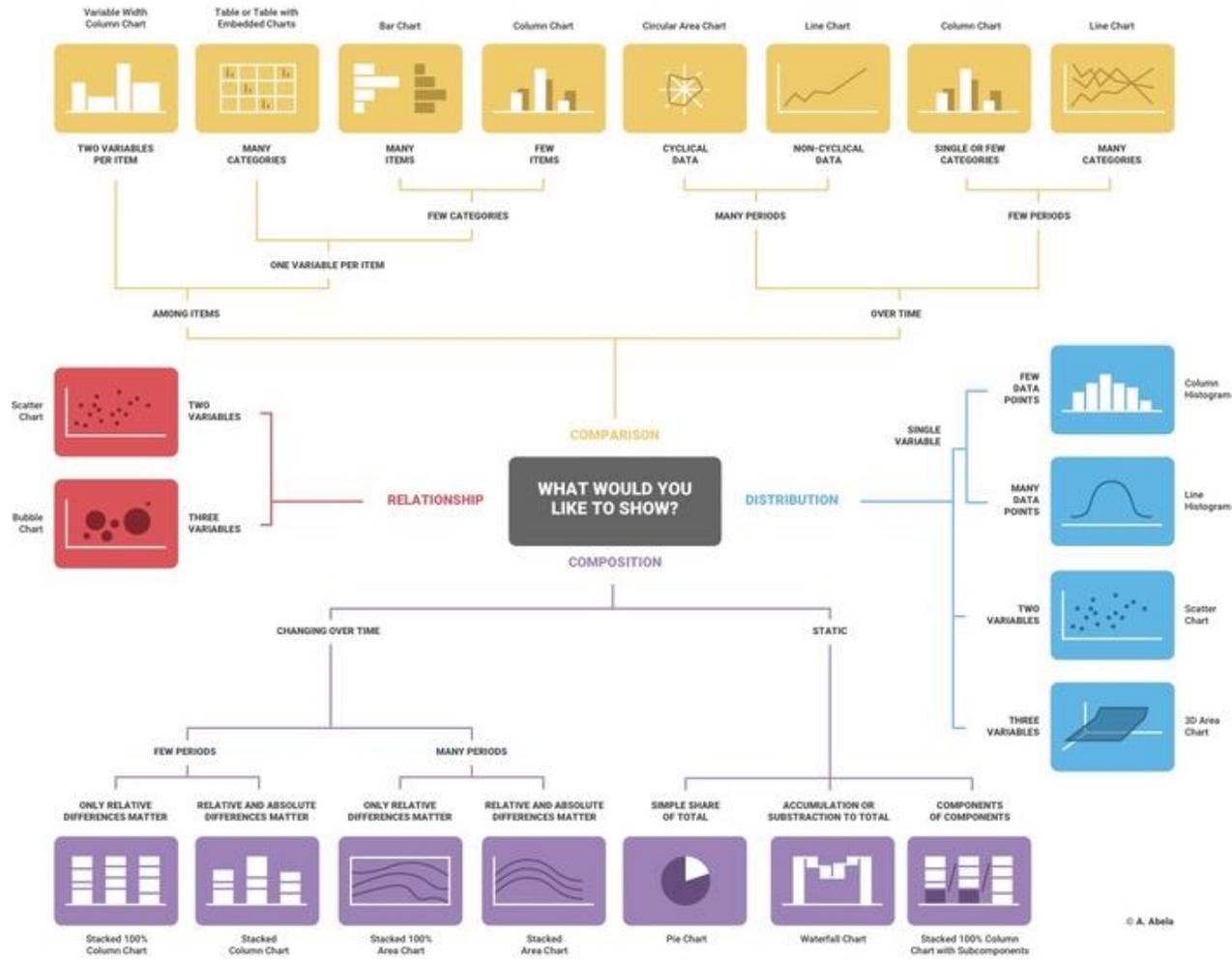
What kind of data do you have? Pick the main type using the buttons below.
Then let the decision tree guide you toward your graphic possibilities.

Numeric | Categoric | Num & Cat | Maps | Network | Time series

https://www.data-to-viz.com/
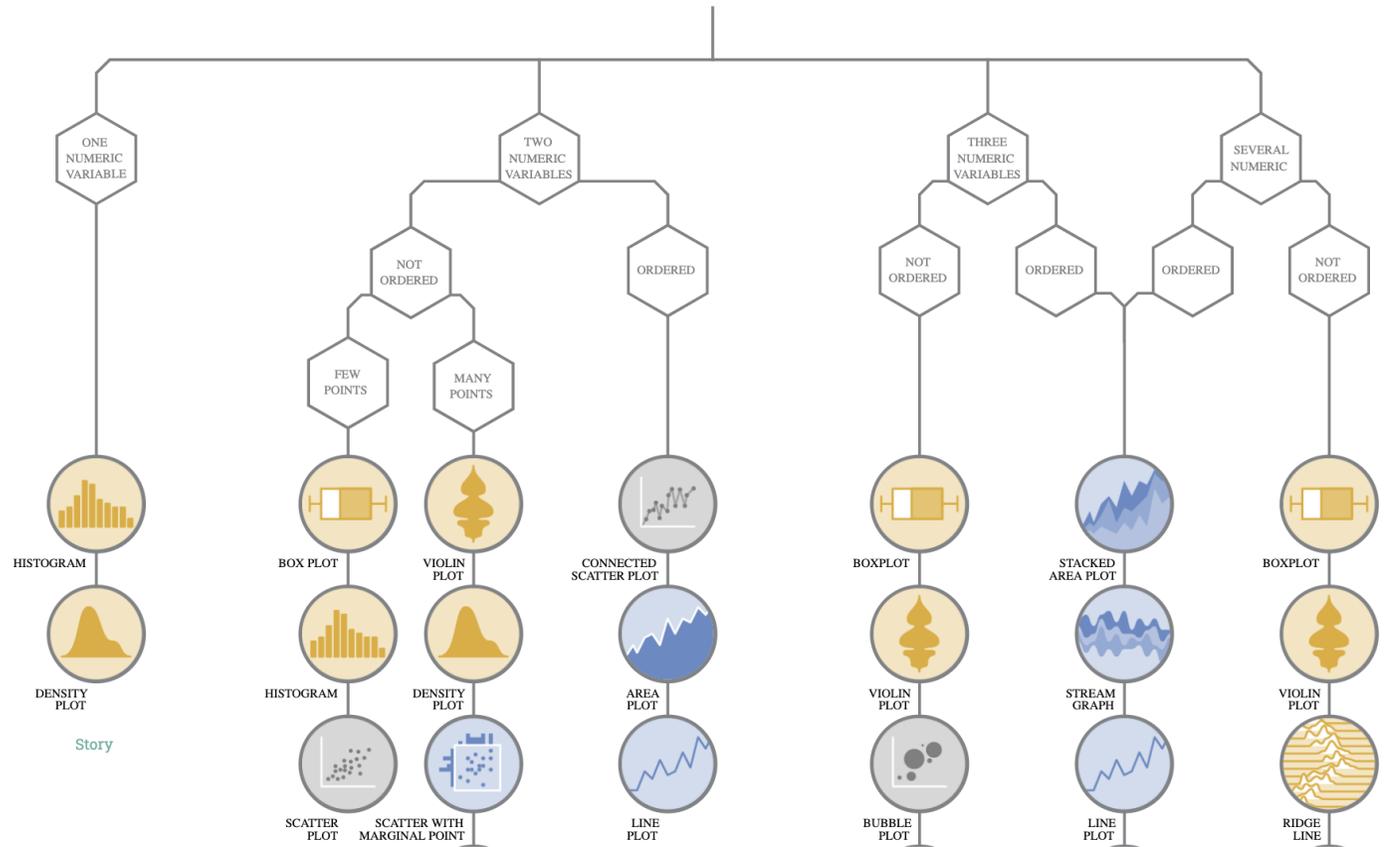
COMPARISON

RELATIONSHIP

What would you
like to show?

DISTRIBUTION

COMPOSITION

Comparing values

Bar charts

- **What is it?** The value for each entity in a category is shown as a bar, where the length corresponds to the value

- **Input data:** One numeric variable for 1 or more categories

- **Why use it?** To compare amounts
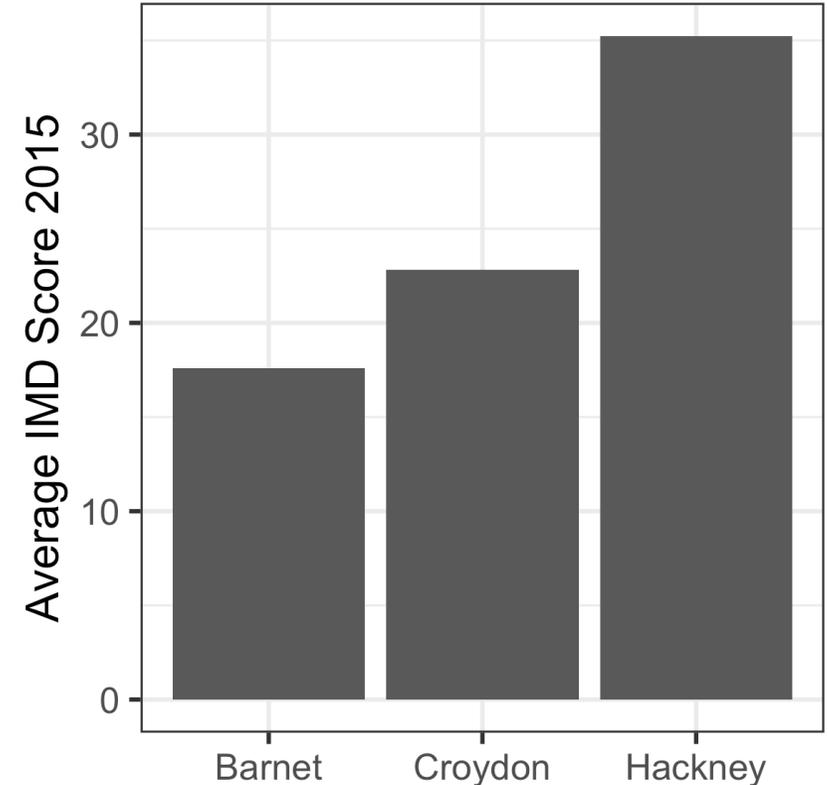
Bar charts

Region	Percent
London and south-east	30.422247
South-west	11.700077
Scotland	9.747461
East of England	9.023659
North-west	8.396877
East Midlands	7.434044
West Midlands	7.311723
Yorkshire and the Humber	6.161340
Wales	4.100153
North-east	3.159430
Northern Ireland	2.542988

Length corresponds to value

Bar chart: Voluntary sector employees by region, June 2019 (%)

Bar chart tips

Voluntary sector employees by region, June 2019 (%)
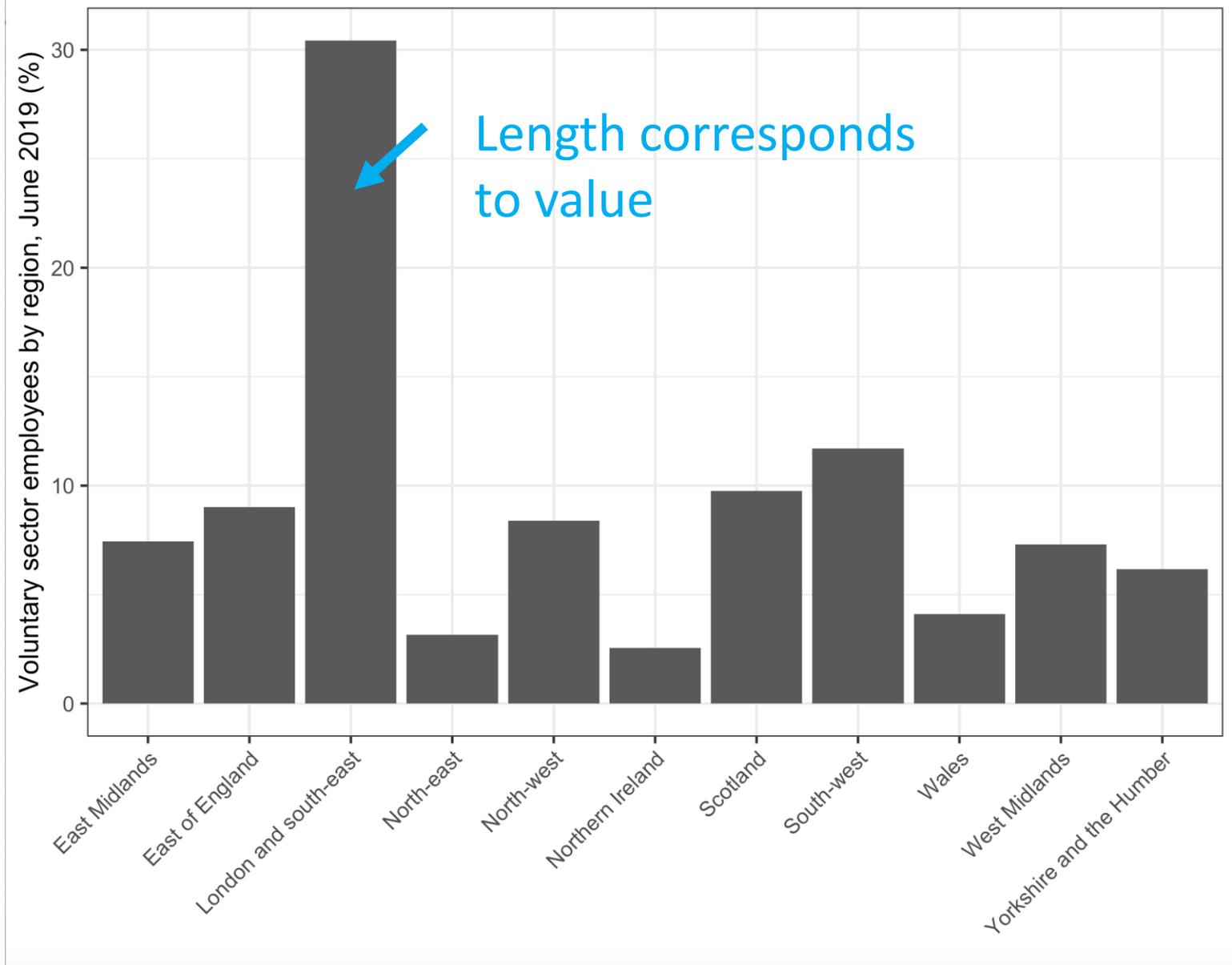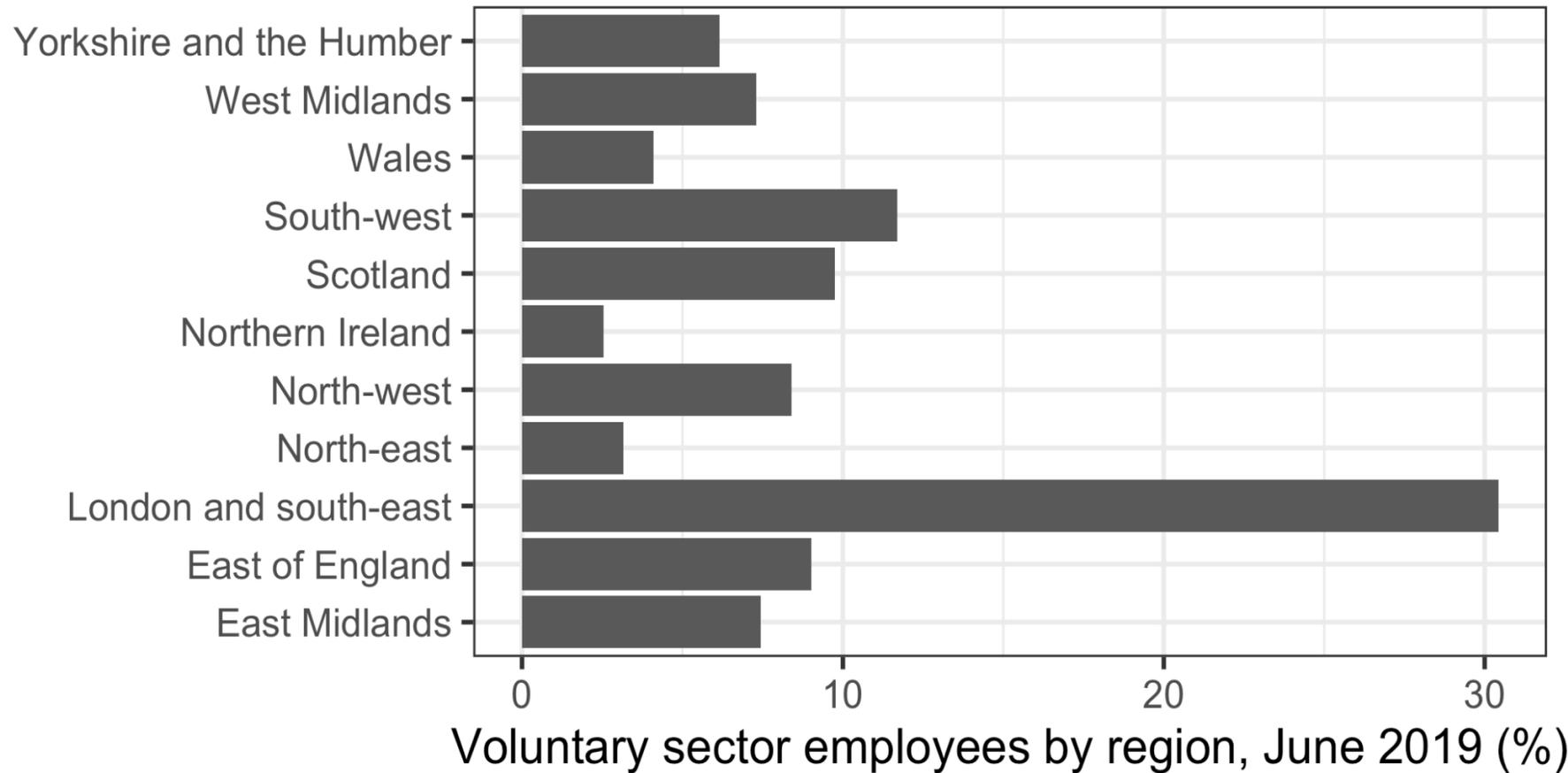
Flip charts with long labels sideways

Bar chart tips

Reorder to enable comparison (only if there is no natural order)

Voluntary sector employees by region, June 2019 (%)

Regions (top to bottom): London and south-east, South-west, Scotland, East of England, North-west, East Midlands, West Midlands, Yorkshire and the Humber, Wales, North-east, Northern Ireland

Bar chart tips

Voluntary sector employees by region, June 2019 (%)

Region	Value
London and south-east	30.4
South-west	11.7
Scotland	9.7
East of England	9
North-west	8.4
East Midlands	7.4
West Midlands	7.3
Yorkshire and the Humber	6.2
Wales	4.1
North-east	3.2
Northern Ireland	2.5

Add value to bar if precise value is important

Spider Chart/Radar Chart

- **What is it?** A 2D chart that shows a series of values for *multiple variables*

- **Input data:** One number for >3 variables

- **Why use it?** To compare values across multiple domains
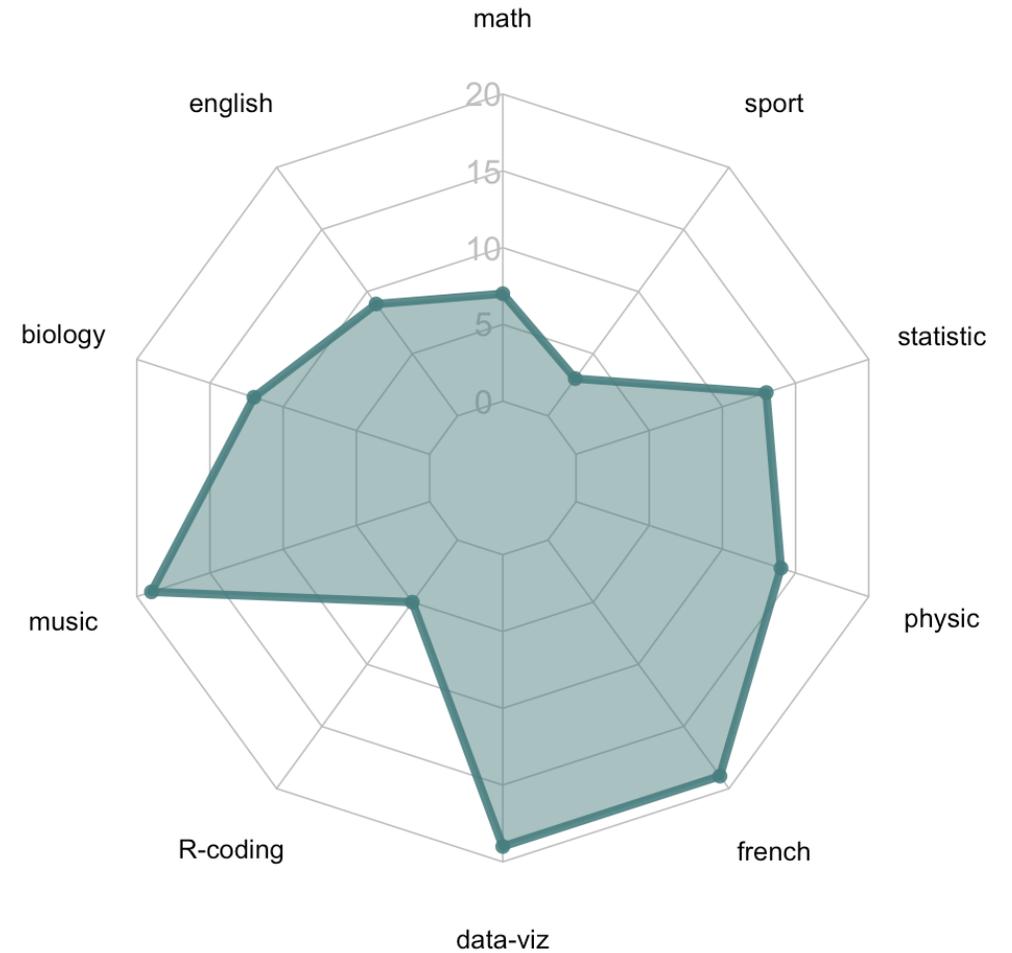
Outcome stars are a type of radar chart
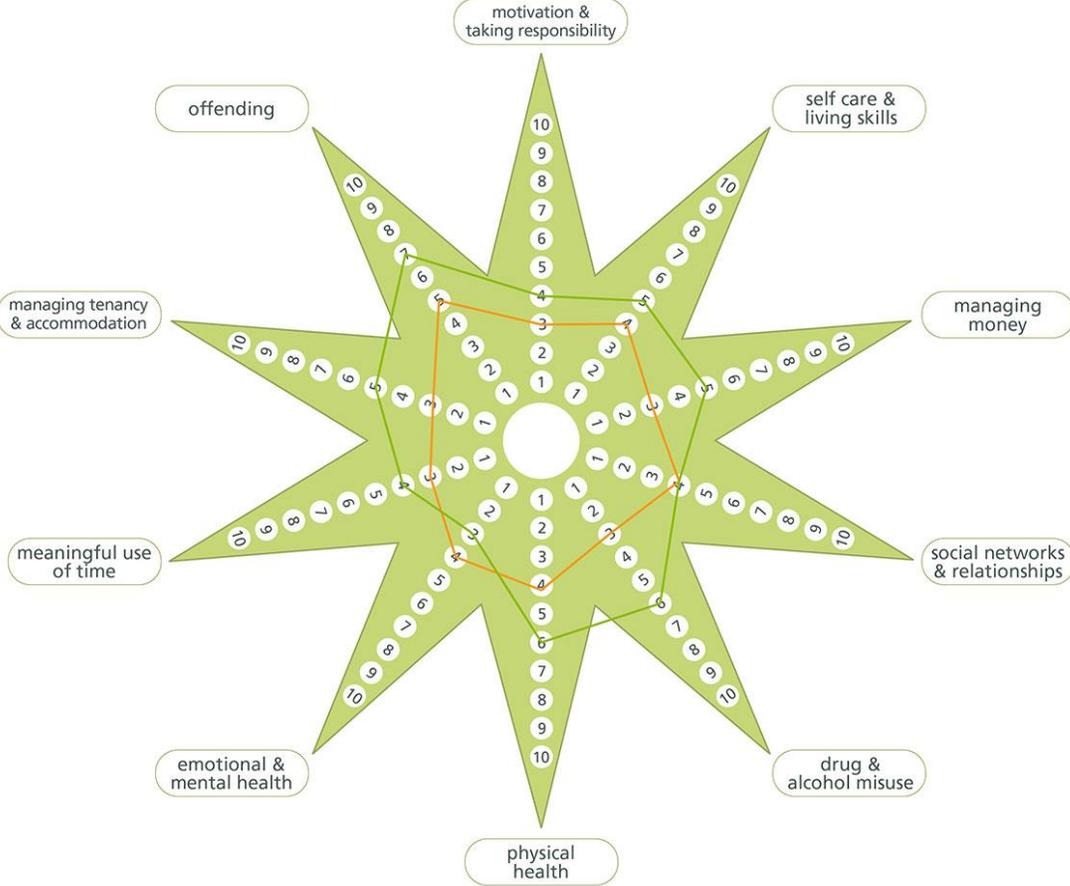
Spider Chart/Radar Chart

Advantages:

- Useful for individuals to compare scores across similarly measured
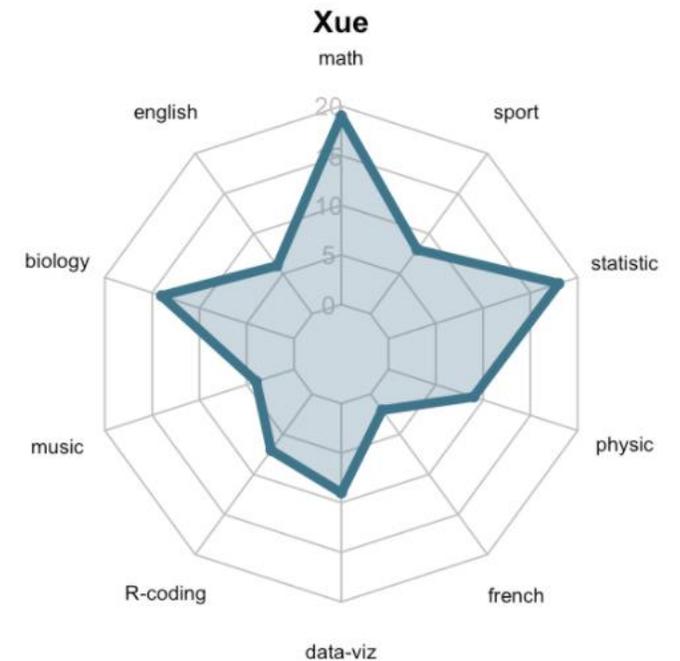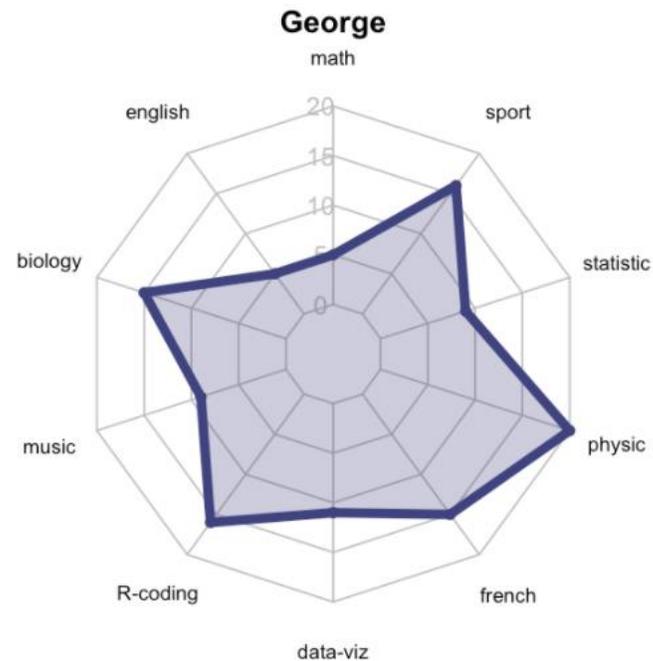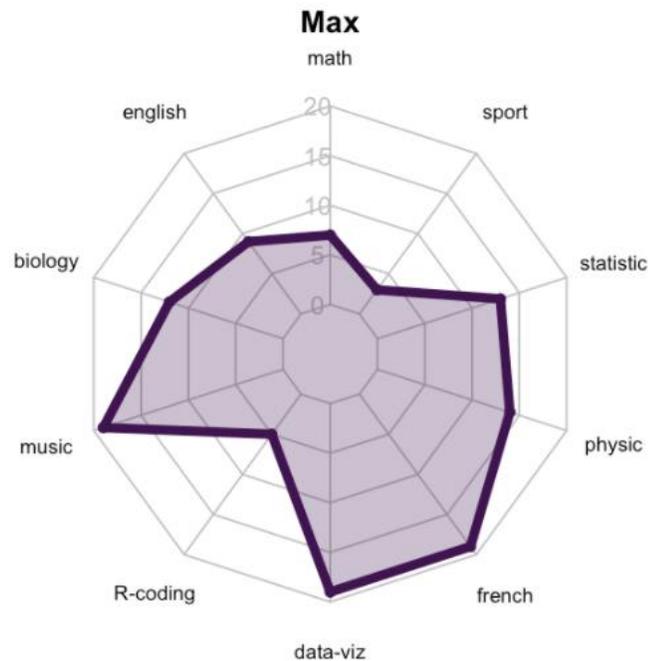
Spider Chart/Radar Chart

Disadvantages

- Circular layout is hard to read

- Hard to determine ranking

- Category order affects the shape

- Data needs to be on the same scale
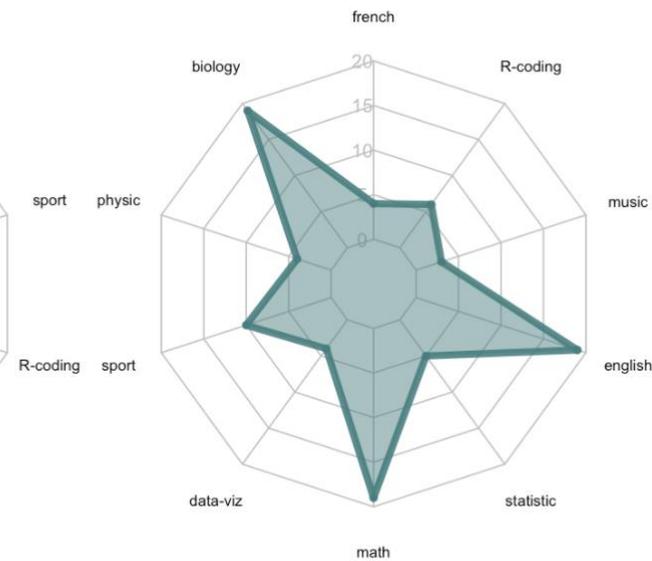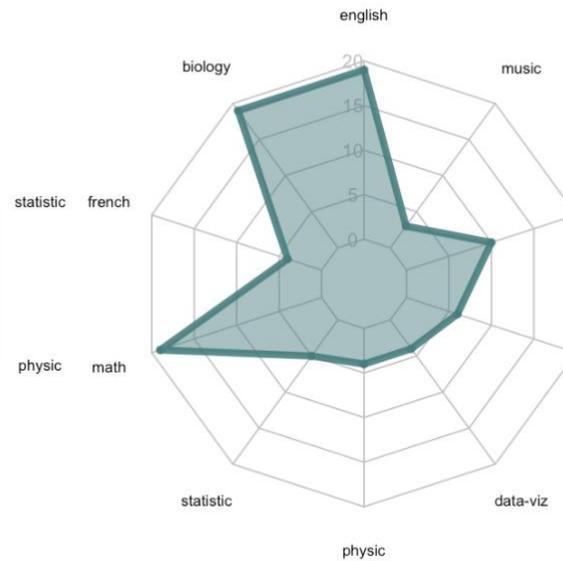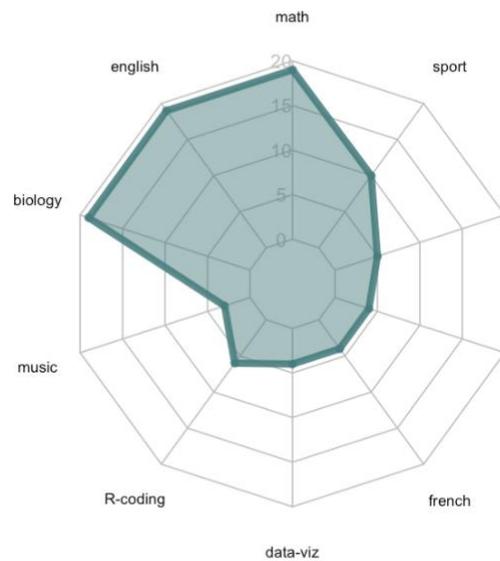
Spider Chart/Radar Chart alternatives
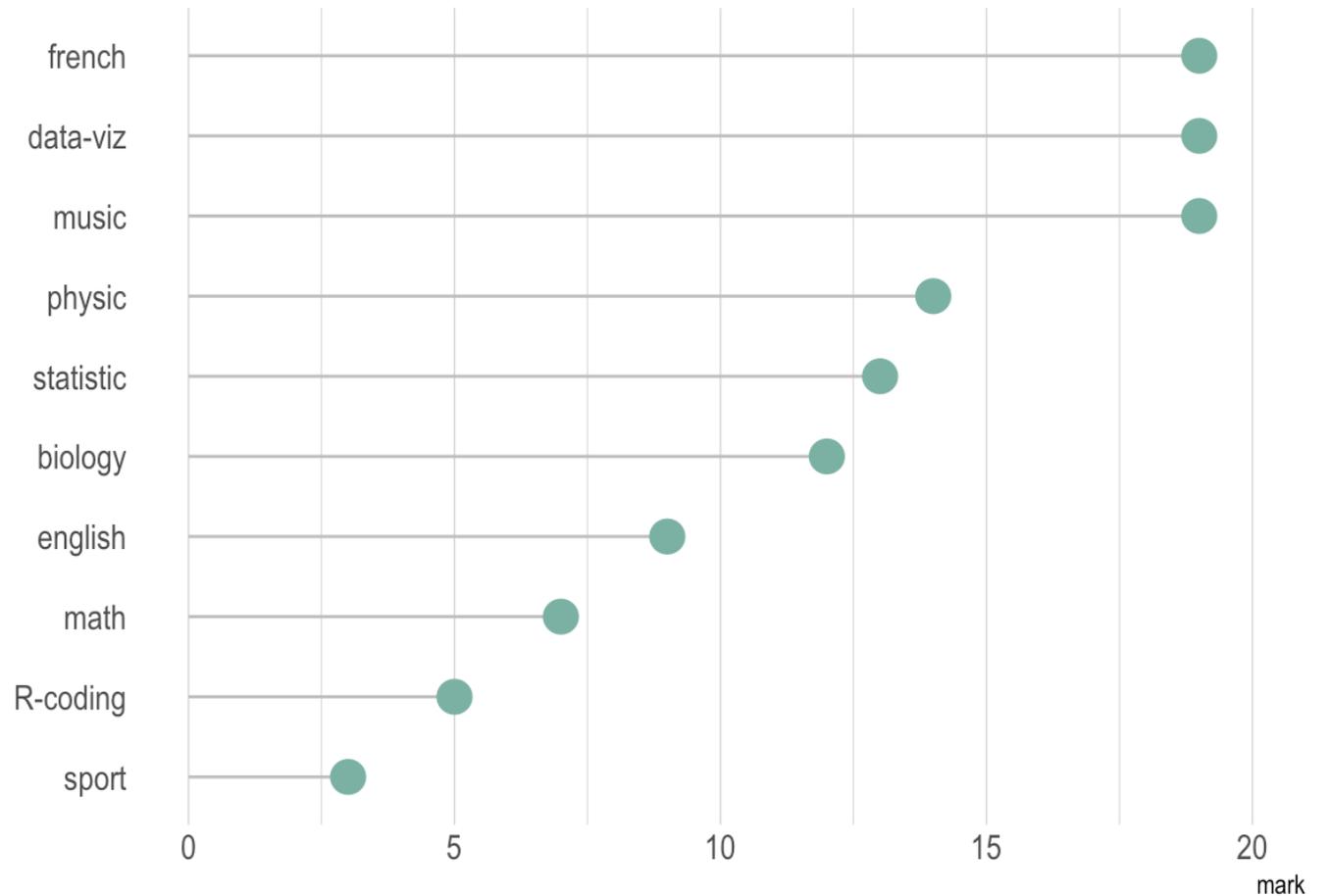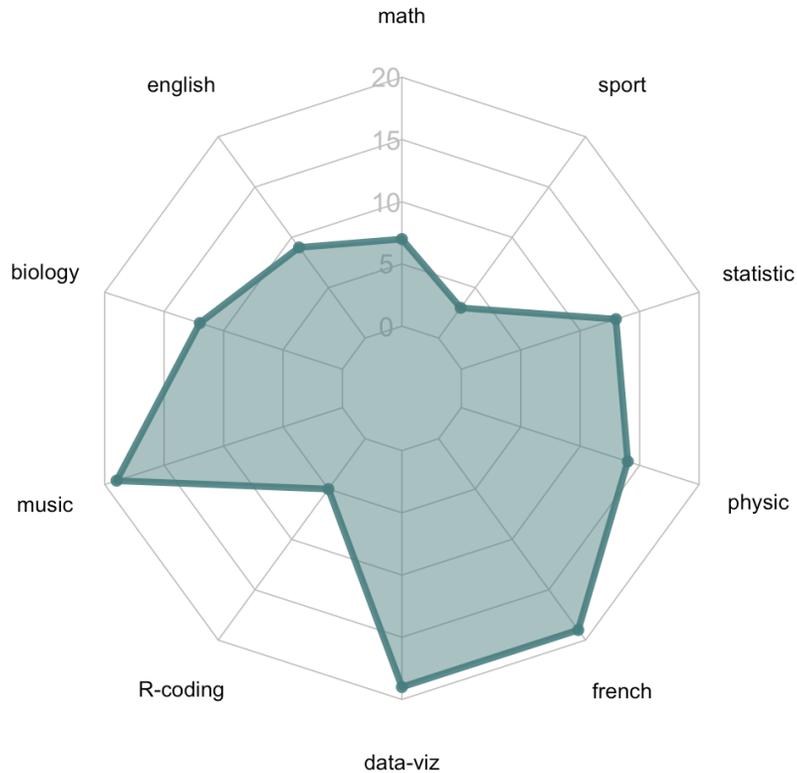
Visualising distributions: understanding variation

Histograms & density plots

What is it?

- shows the *distribution* of numeric data
- A *distribution* shows the probability of different possible values

Input data: one numeric variable

Why use it? To understand the frequency of different values in your data, e.g. if some values are more common than others

Histograms

Input data: Grant funding awarded to different organisations

Amount Awarded
49495
55709
81483
88157
71514
83795
44654
59753
110000
51478

Each row is 1 grant

Within range 50000-75000

BBC Children in Needs Grants

number of organisations

Amount Awarded

Over 1500 orgs receive £25-50k

Bin width = 25000

Few receive >£150k
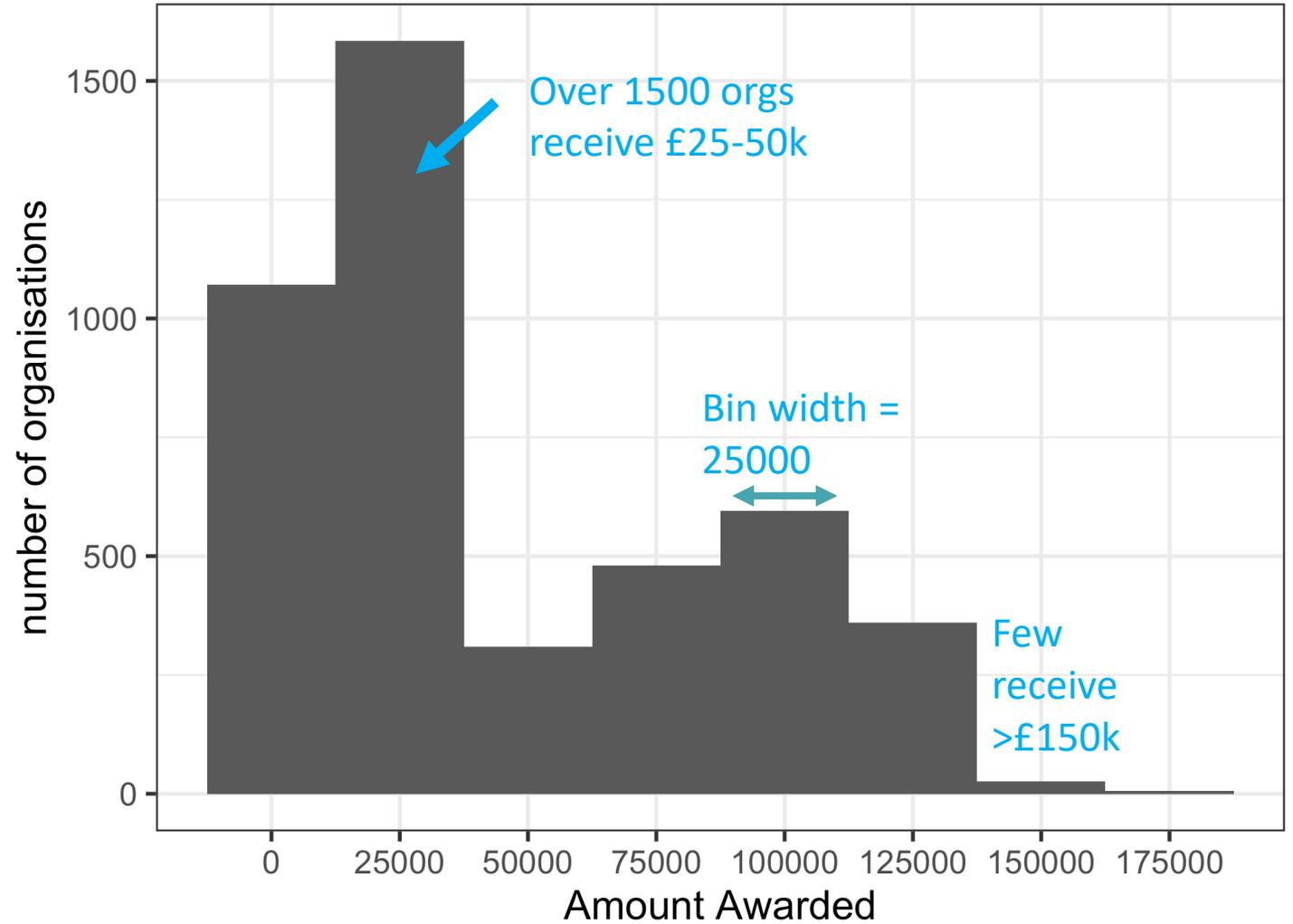
Common pitfalls: bin width

BBC Children in Needs Grants
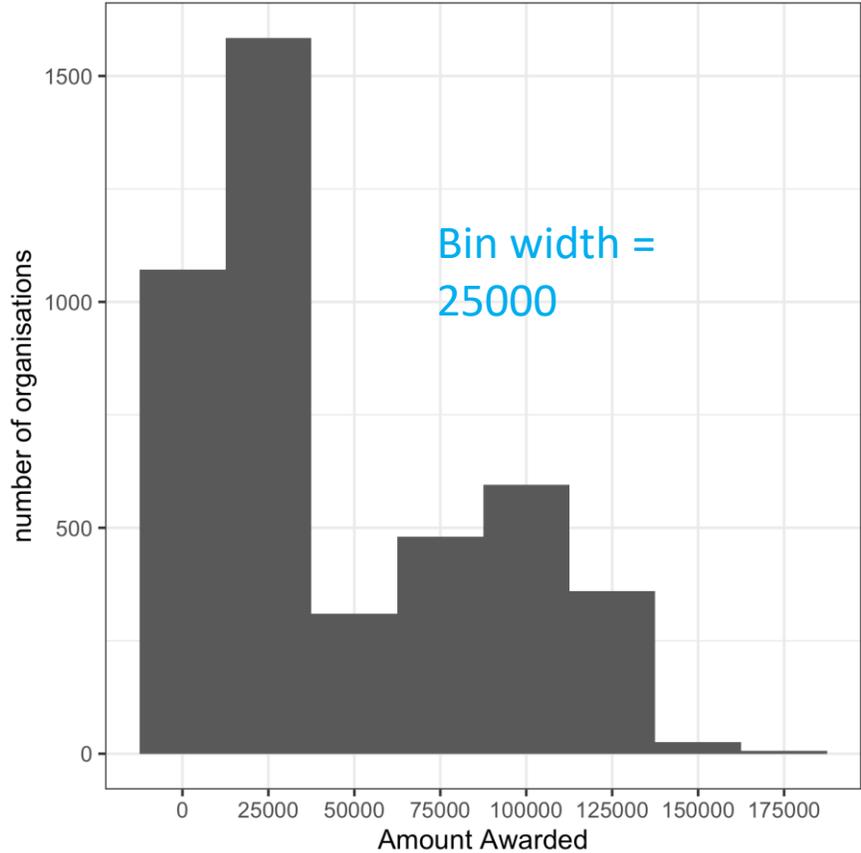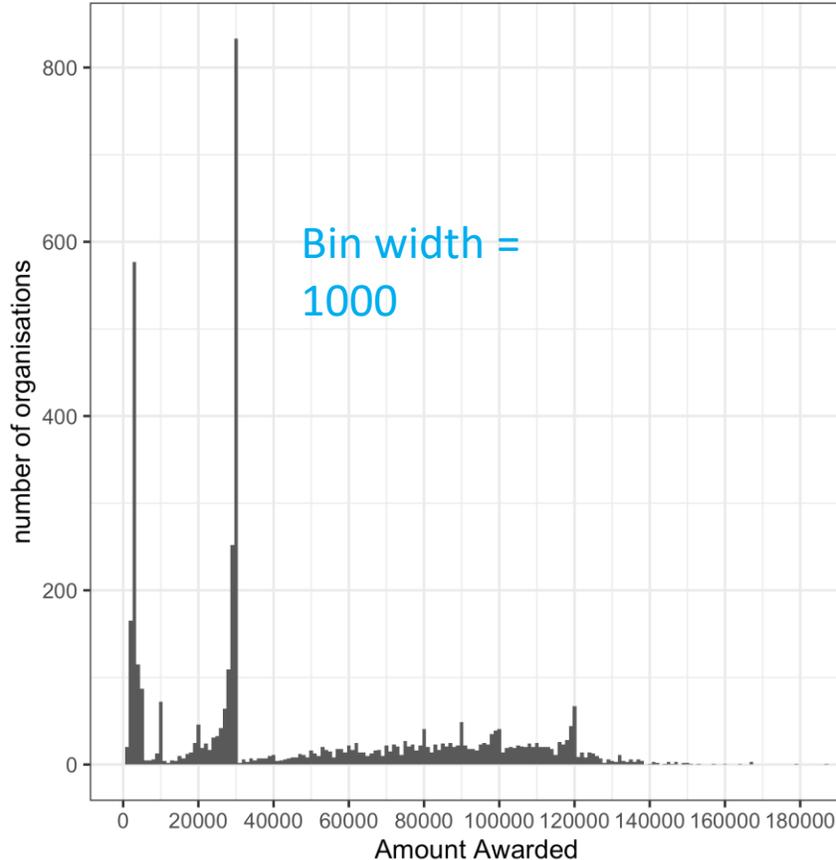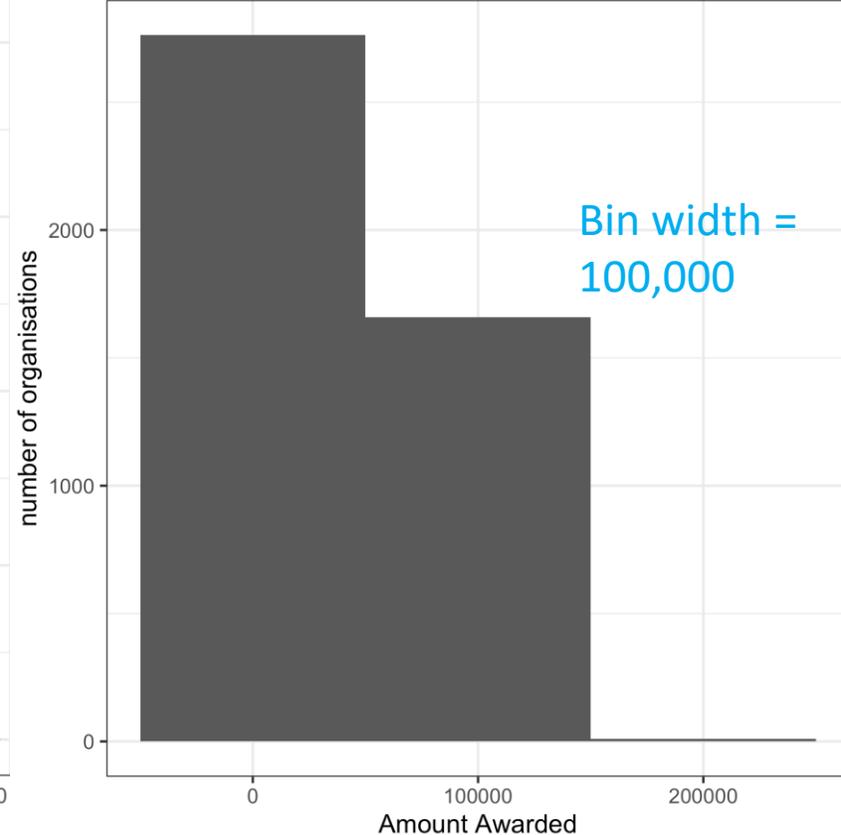
Bin width = 25000

BBC Children in Needs Grants

Bin width = 1000

BBC Children in Needs Grants

Bin width = 100,000

Understanding composition

Pie charts and donut charts

- **What is it?**

  - A circle divided into sections where each piece of the pie represents the proportion

  - Donut chart uses ring vs circle

- **Input data:** Proportion/Percentage for each value in a category

- **Why use it?** To compare proportions

FRACTION OF THIS IMAGE WHICH IS WHITE

FRACTION OF THIS IMAGE WHICH IS BLACK

Xkcd.com
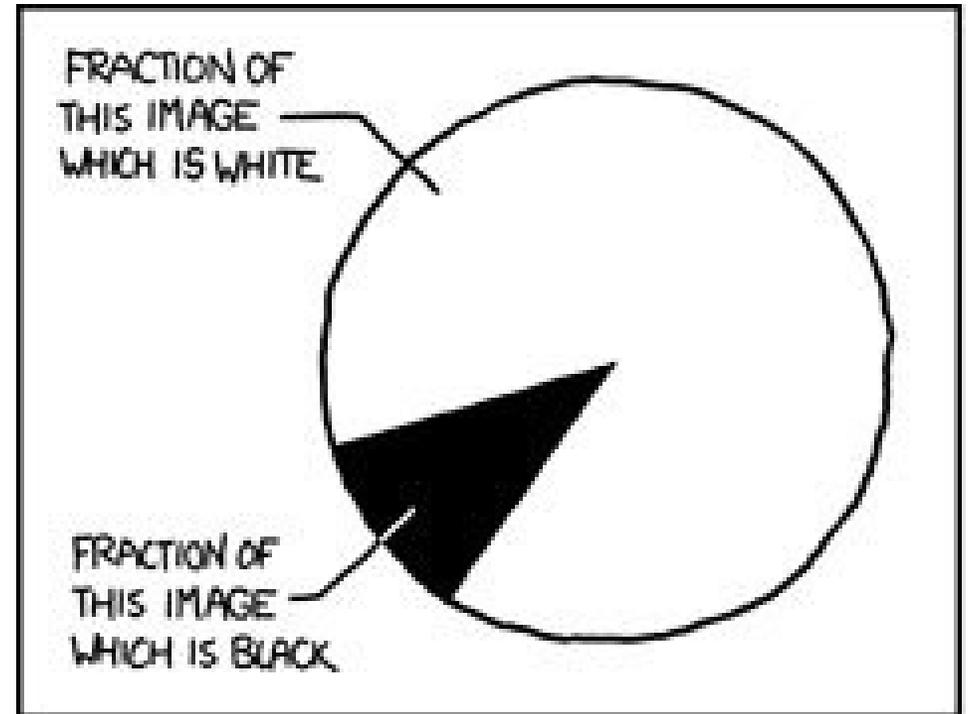
Pie charts and donut charts

In the past 12 months, have you wanted counseling from a mental health professional?

No
16%

Yes, but I didn't get it
46%

Yes, and I got it
38%
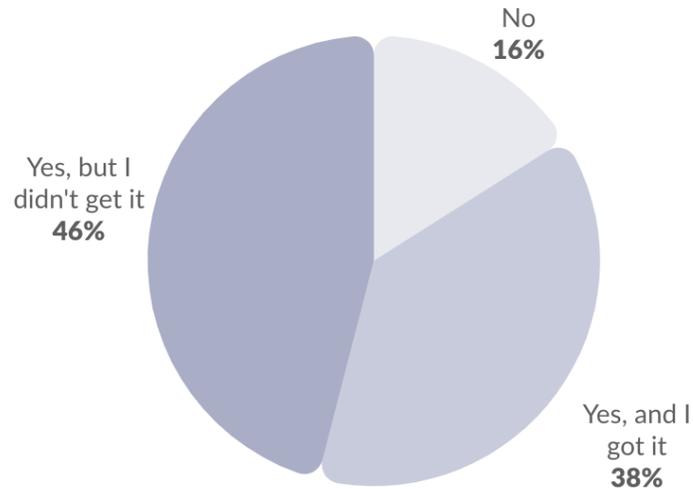
How much of our support went to organisations led by Black, Asian and minority ethnic communities?

We wanted to show what proportion of these grants went to organisations which self-defined as "BAME-led" when they applied for funding.

9%

49 out of 545 Fast Response Grants went to BAME-led organisations.

91%

- Best used with small datasets

- Best used with simple fractions (1/10, ½, ¼)

Pie charts – what does this tell you?

Pie chart alternatives: bar charts
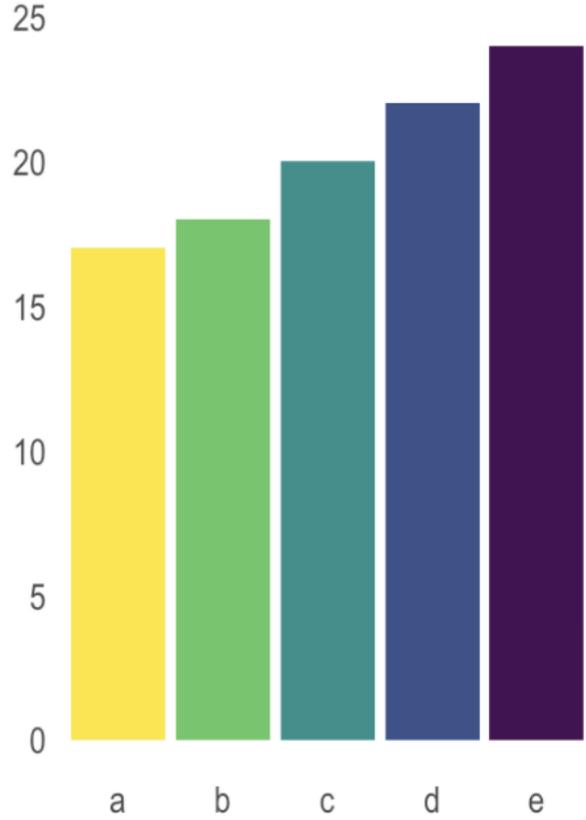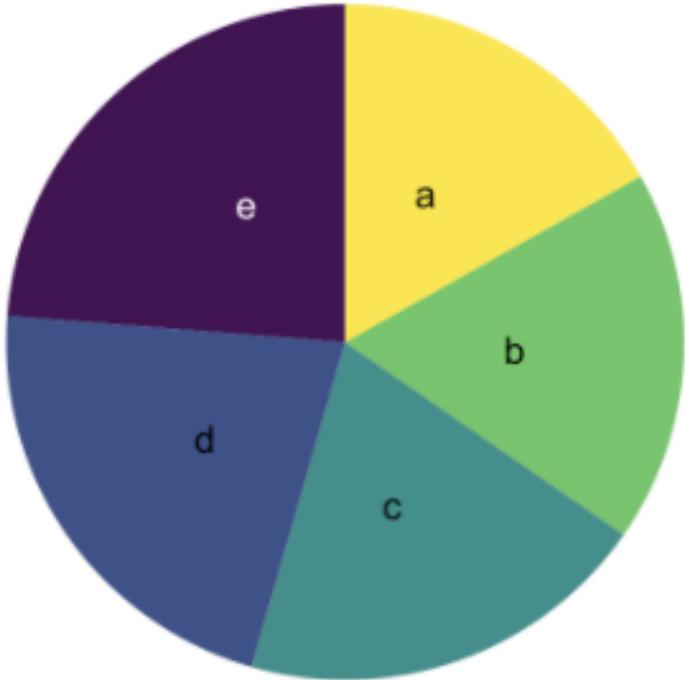
Pie chart alternatives: stacked bar charts

Micro and small organisations receive a larger proportion of income from the public than other organisations

Breakdown of income by source and size of organisation, 2017/18 (%)

■ The public ■ Government ■ National Lottery ■ Voluntary sector ■ Private sector ■ Investment

	The public	Government	National Lottery	Voluntary sector	Private sector	Investment
Micro and small	58	13	8		17	
Medium	49	23	10		4	12
Large	49	29	8		5	8
Major	45	36	7		7	4
Super-major	46	27	16		4	7
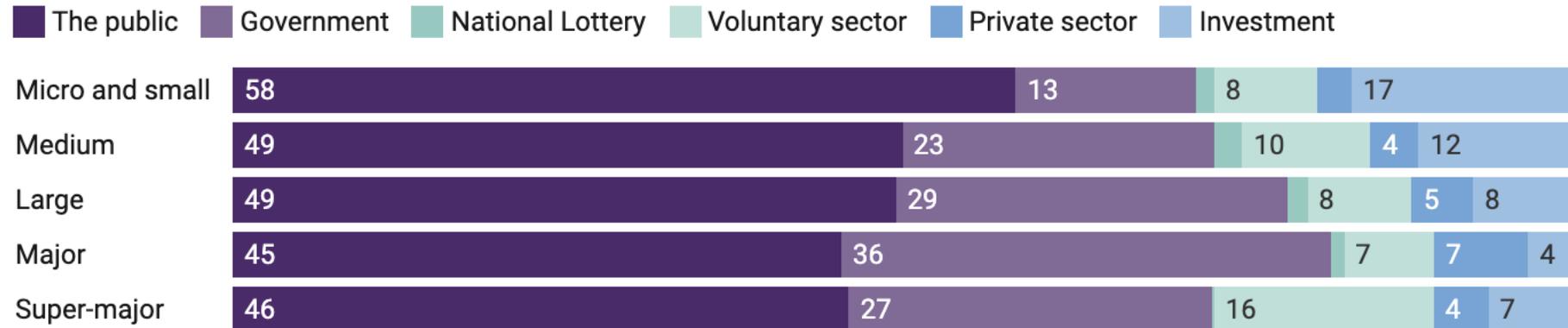
Source: NCVO, Charity Commission • Get the data • Created with Datawrapper

Pie chart alternatives: pictorial fraction chart

LGBT representation by job level

Sexual orientation of staff at different job levels, National Portfolio Organisations: All paid staff (2018-19)

● LGBT ● Straight ● Prefer not to say ● Not known

Artists

LGBT	7%
Straight	37%
Prefer not to say	12%
Not known	44%

Managers

LGBT	9%
Straight	65%
Prefer not to say	11%
Not known	14%

Specialist

LGBT	8%
Straight	50%
Prefer not to say	10%
Not known	32%

Other

LGBT	7%
Straight	47%
Prefer not to say	14%
Not known	31%
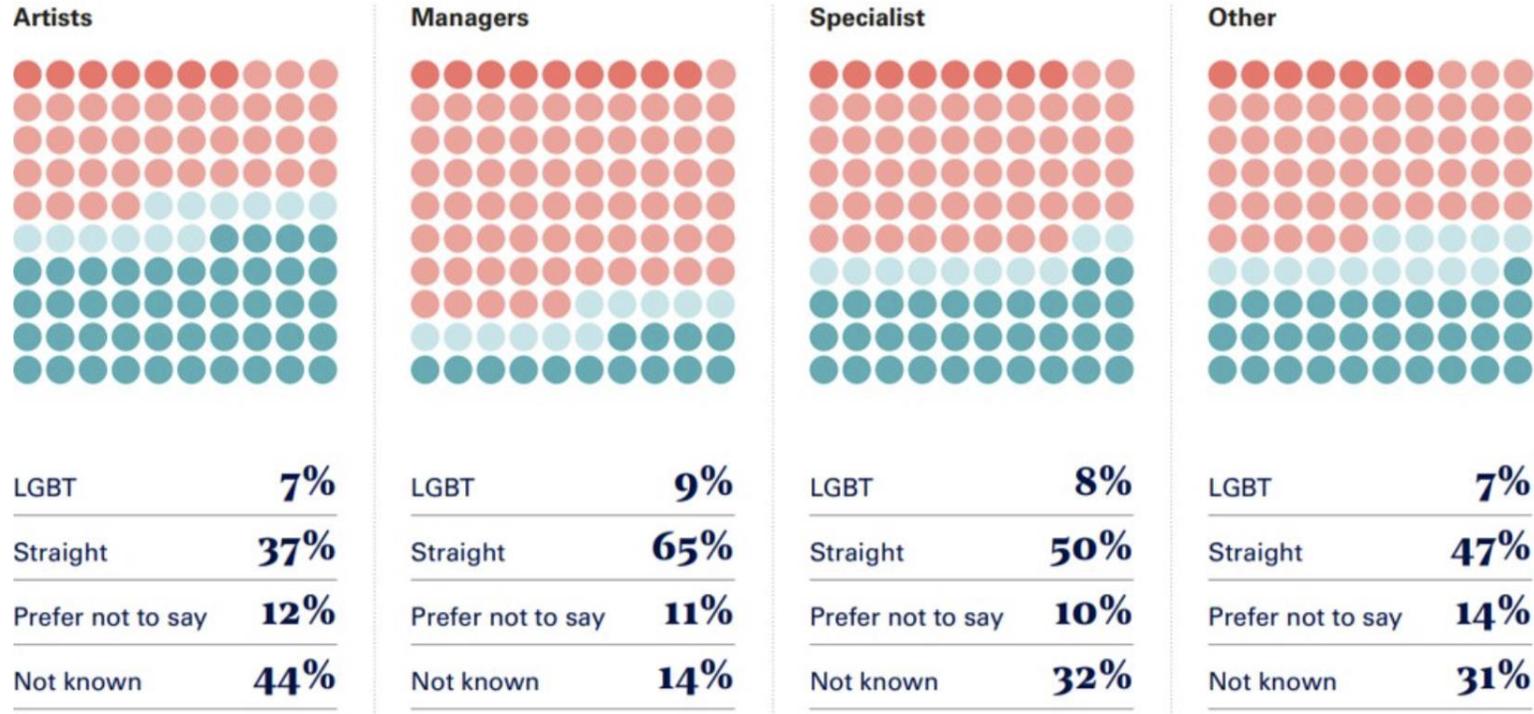
COMPARISON

RELATIONSHIP

What would you like to show?

DISTRIBUTION

COMPOSITION

Understanding relationships

Scatterplots, line charts, & area charts

- **Scatterplots:** show how one factor (variable) changes as another changes, e.g. weight and height
- **Input data**: Two numeric variables
- **Why use it?**
  - To understand relationships between two factors

Scatterplots

income_deprivation	child_poverty
3036	1058
2891	1057
3130	1082
2377	822
2061	640
2931	1036
3536	1366
2924	997
3039	1041
1683	493
2718	955
2485	830
2613	929

Each row is data from one ward in London
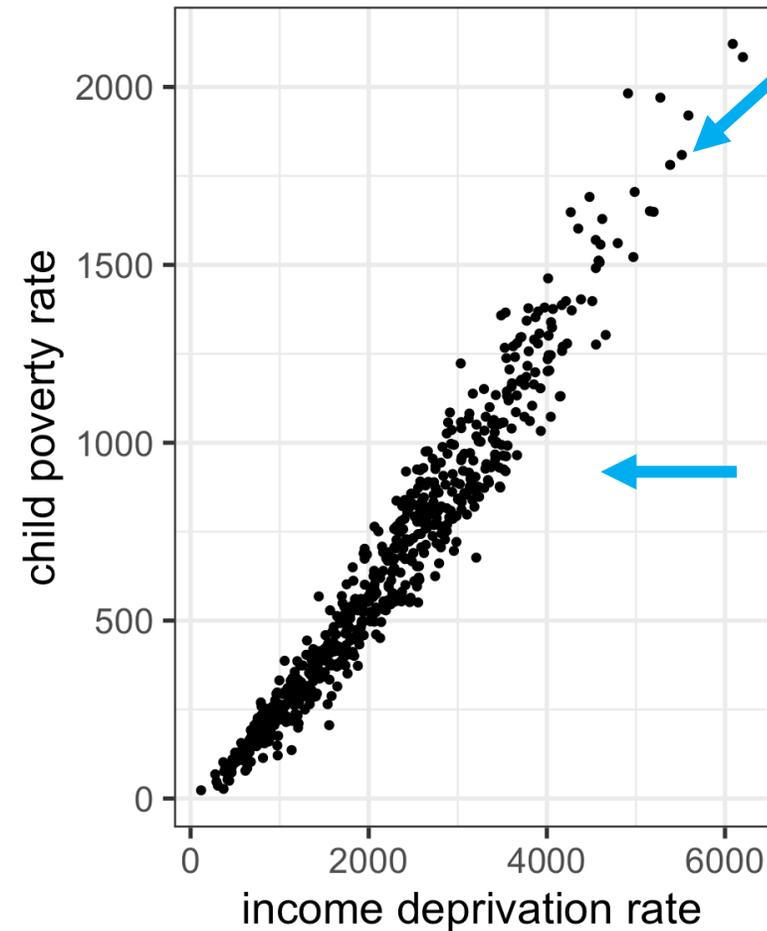
Each point is one row or sample

Overall pattern can tell you the type of relationship

Common pitfalls: overlapping points
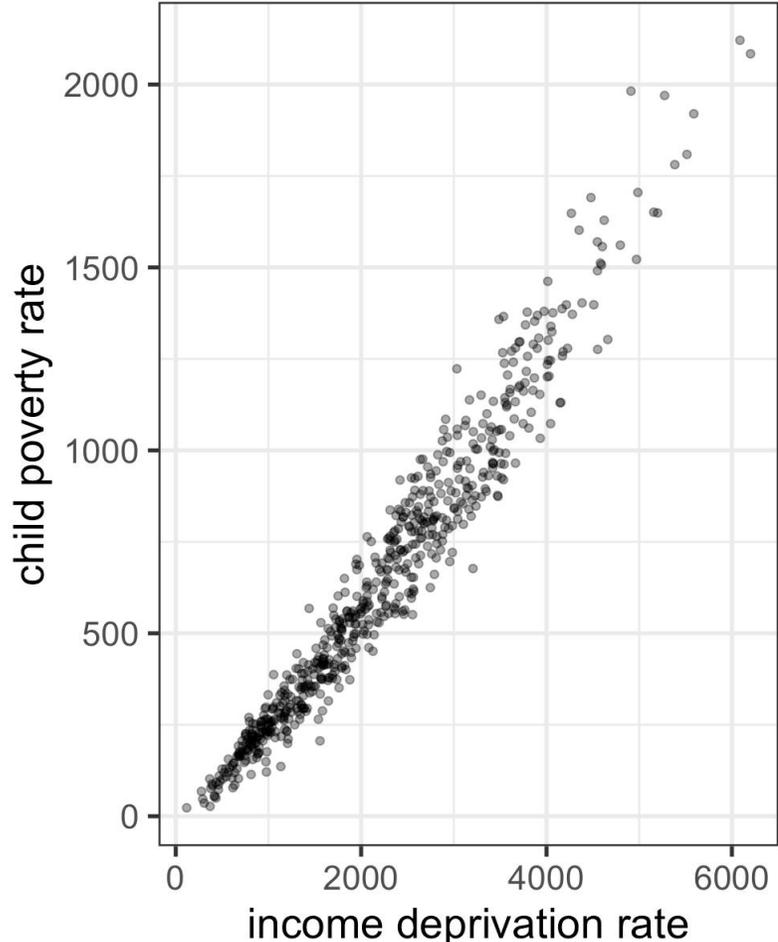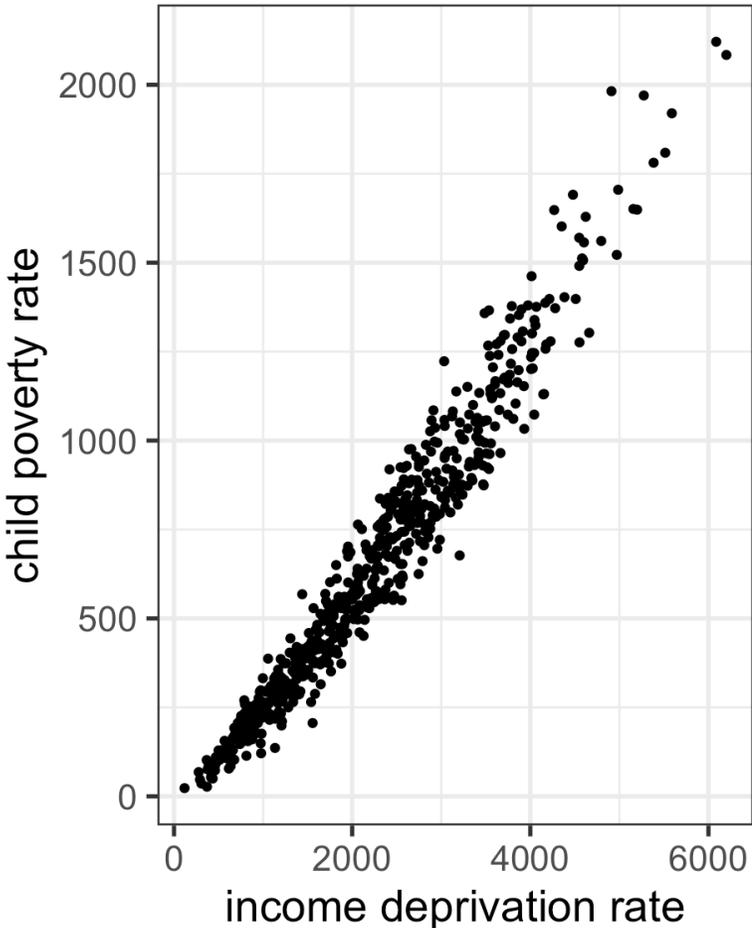
Scatterplots

On average, places that were hit hardest in the spring are suffering the most in the autumn

Weekly cases per 100,000 people*, spring peak vs autumn peak. Each circle represents one region

France

Autumn peak

Paris
Hauts de France
Grand Est

Italy

Liguria   Trento
Bolzano
Aosta
Lombardy

Spring peak
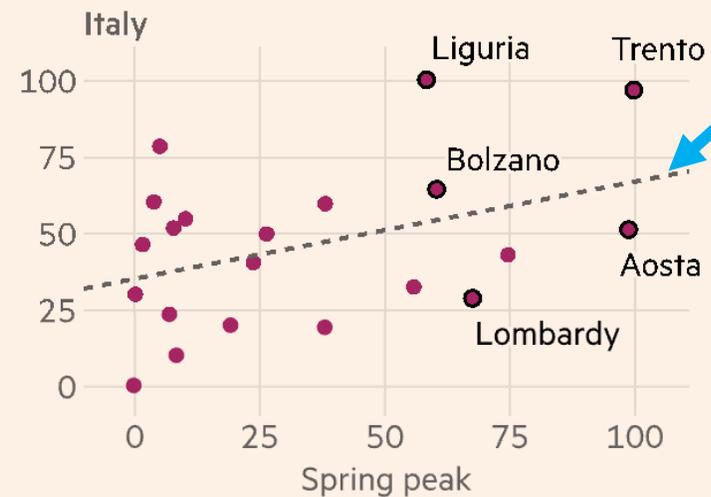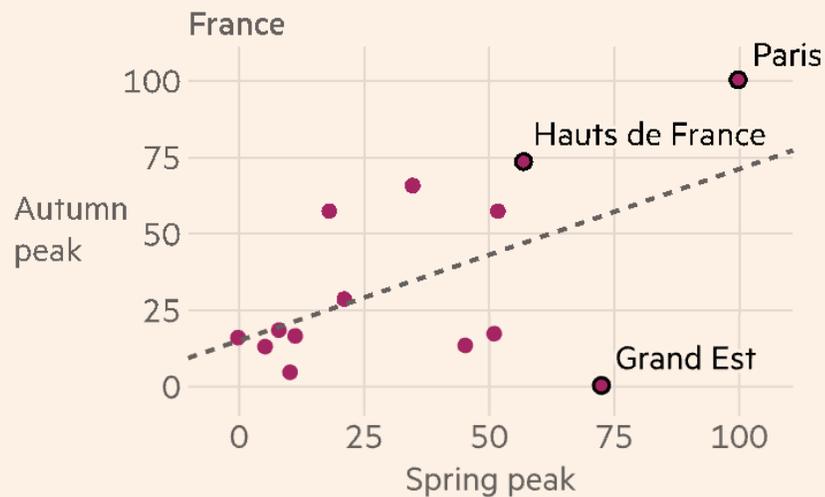
Line of best fit can help you compare trajectories

Line charts

The number of people employed in the voluntary sector has risen steadily

Women — Men — Total

Source: NCVO/Labour Force Survey

BBC
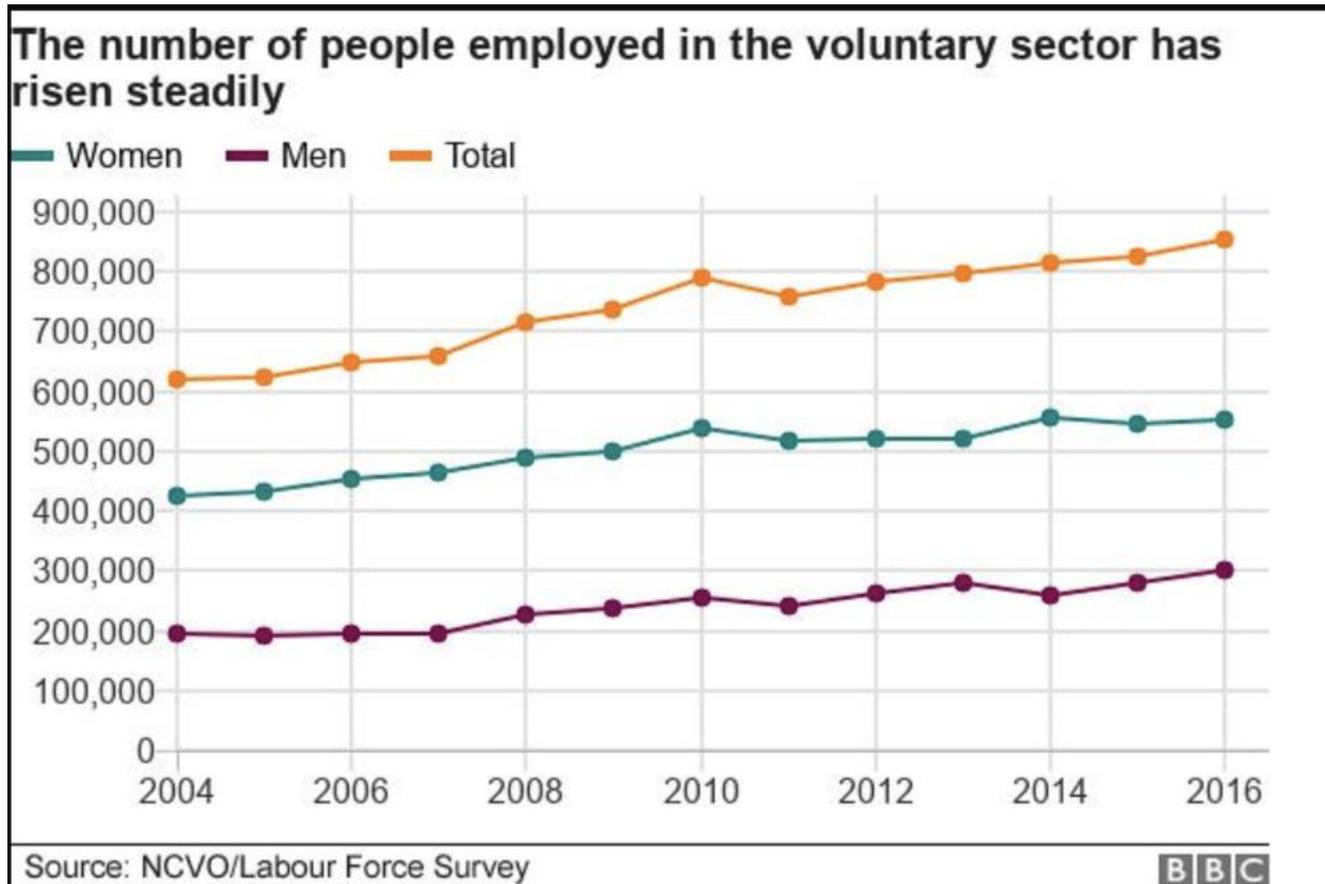
- Like a scatterplot, but the markers are connected by a line
- The x-axis (horizontal axis) values are ordered, often over time

Common pitfalls: Spaghetti

Citizens Advice top 25 web pages March 2020
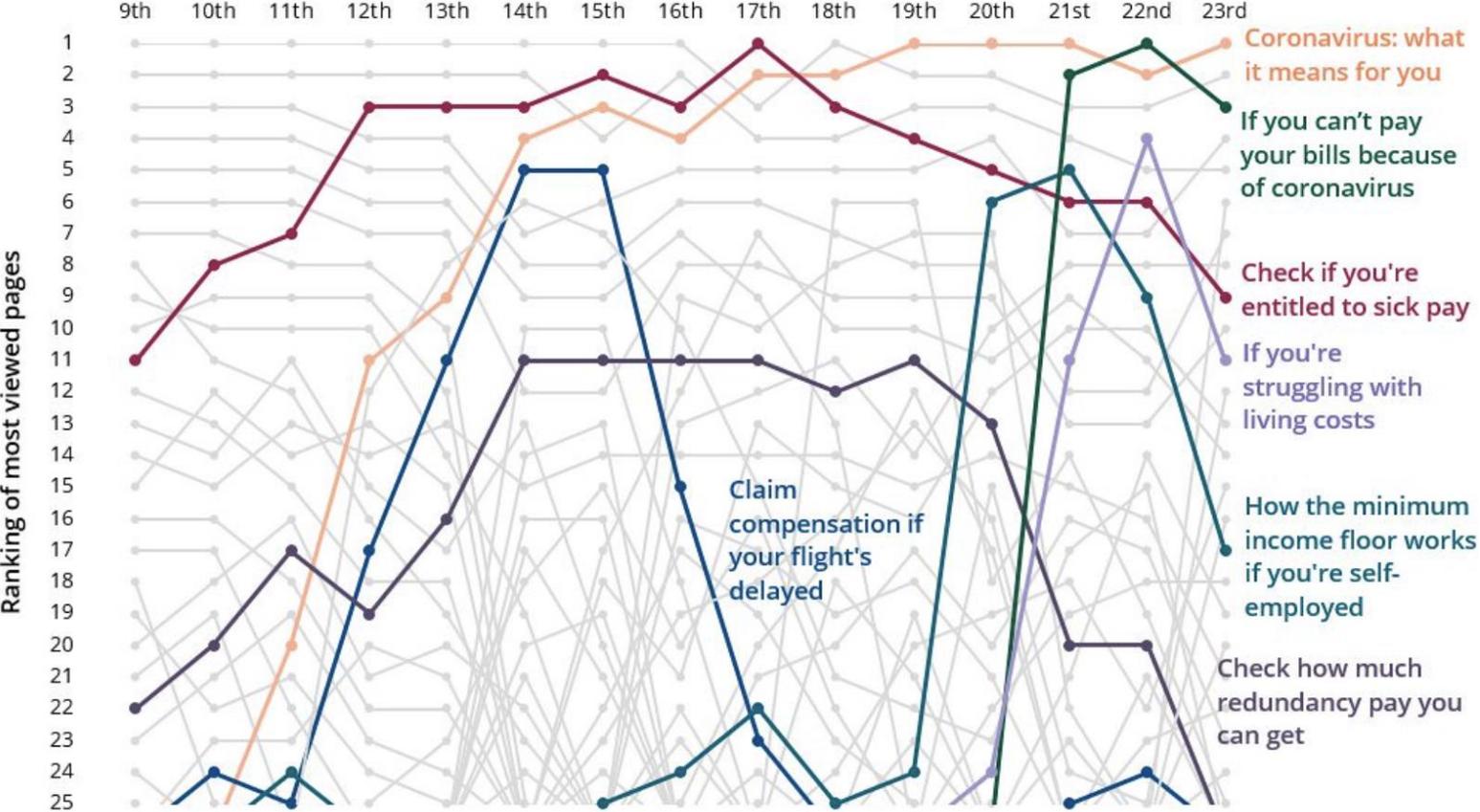
Area charts

COVID−19 cases, per day

700K
600K
500K
USA
400K
300K
200K
SOUTH AMERICA
EUROPE
100K
INDIA
0

Feb 2020 · Mar · Apr · May · Jun · Jul · Aug · Sep · Oct · Nov · Dec · Jan 2021

Seven day rolling average of number of people confirmed to have COVID-19, per day (not including today). This chart gets updated once per day with data by Johns Hopkins. Johns Hopkins university doesn't provide reliable data for March 12 and March 13.

Source: Johns Hopkins CSSE · Get the data · Created with Datawrapper

Like a line chart but the area is filled in to highlight the trend

Questions & Break

Get feedback & Refine

↘ Does every part of the figure have a purpose?
↘ What can you remove?
↘ Do you need additional text for storytelling?
↘ Consider colours
  ★ Use consistently
  ★ Use to highlight important information
  ★ Consider accessibility

Thinking about colour

Avoid colouring for the sake of it
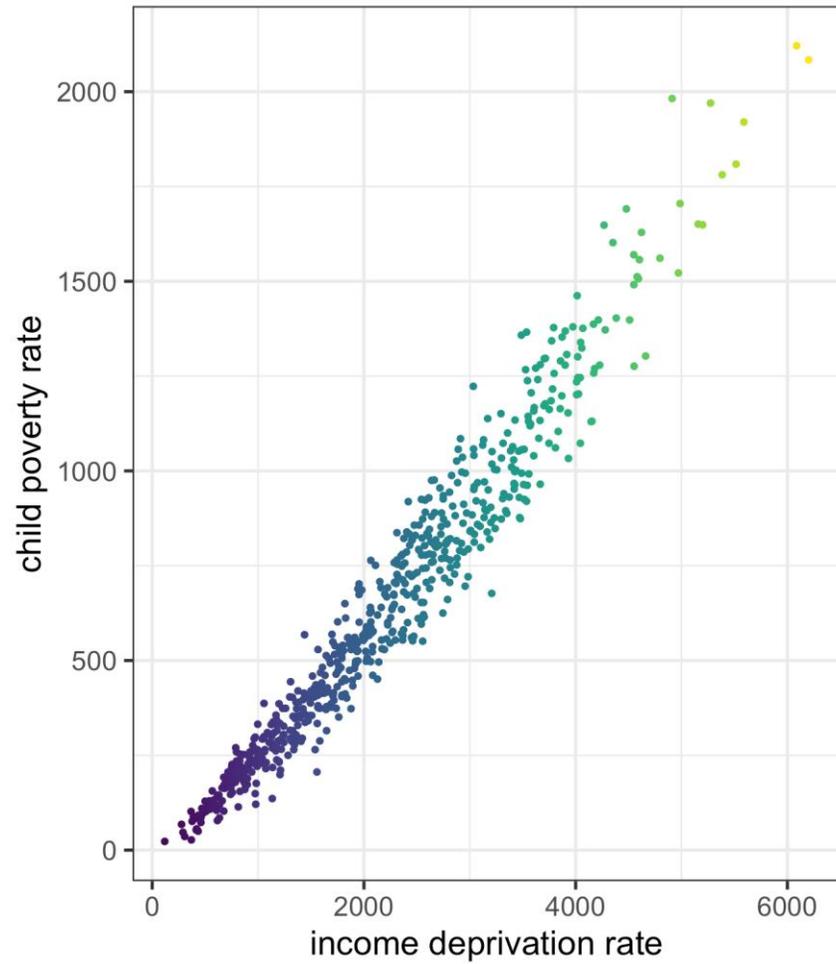
Use colour scales appropriately

Figure 2. Sequential, diverging, and qualitative color schemes (Brewer 1994).

https://cartographicperspectives.org/index.php/journal/article/view/1538/1819
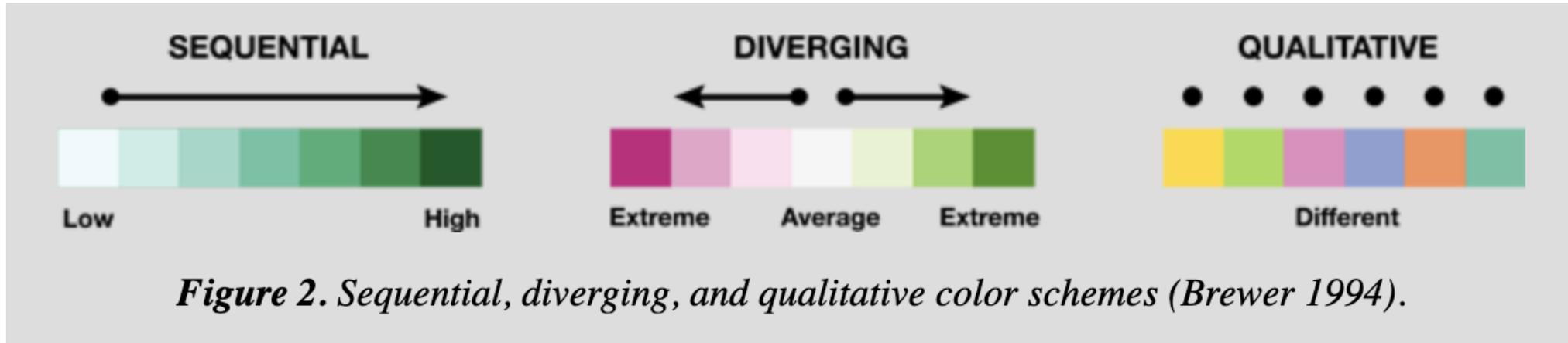
Shots given   Doses distributed   Doses used

Share of state population given at least one dose

14   15   16   17   18   19%

15.6% of the US population given at least one dose
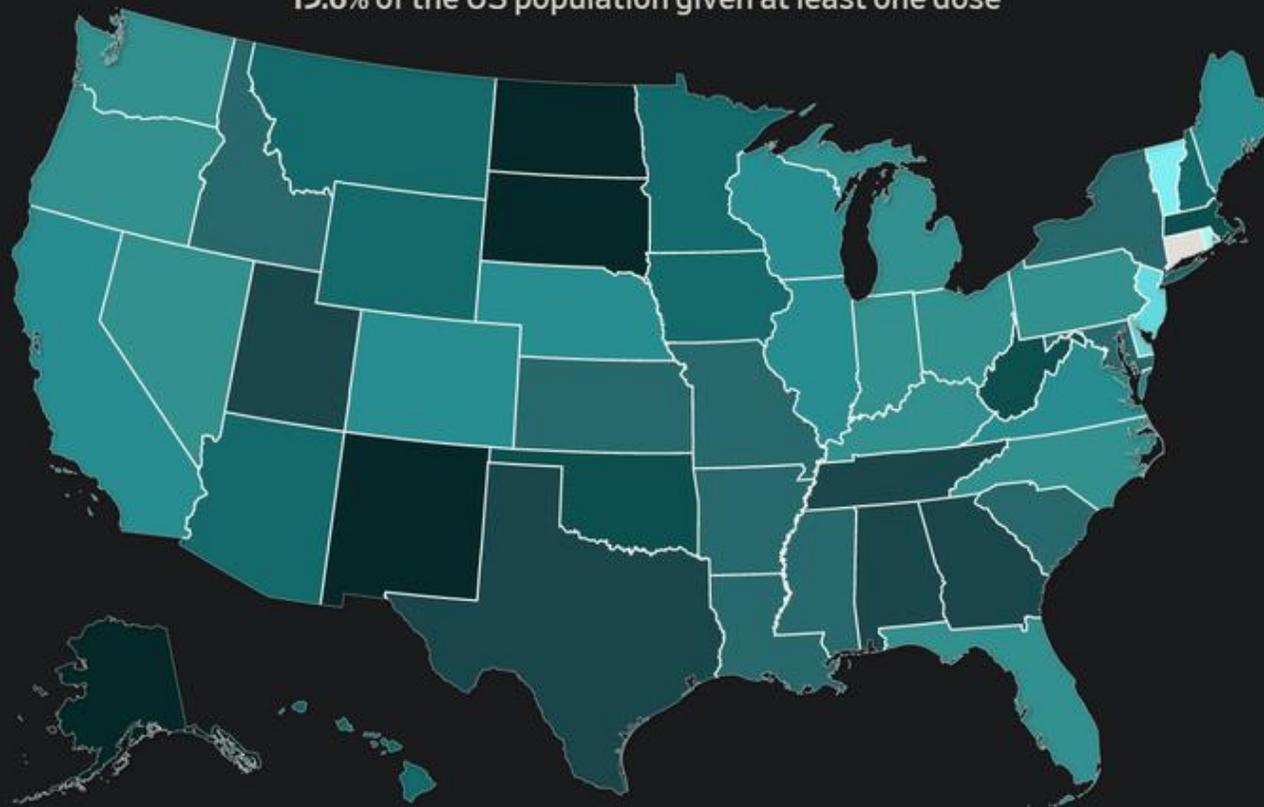
Note: Total includes U.S. territories and federal agencies. Last updated March 2, at 5:00 p.m.
Source: Centers for Disease Control and Prevention

Avoid colouring too many categories
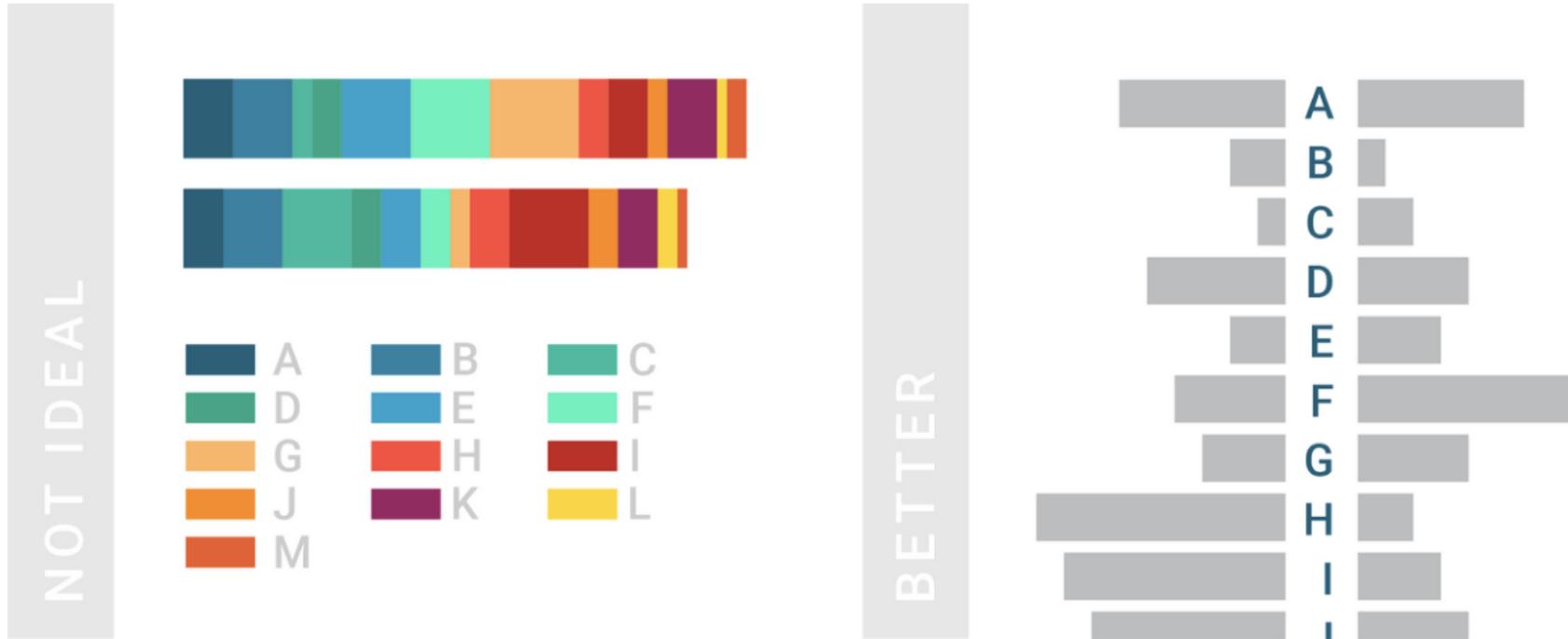
Example of an ideal and non-ideal chart. (Chartable.com, 2018)

Design for colour-vision deficiency

original

deuteranomaly

protanomaly
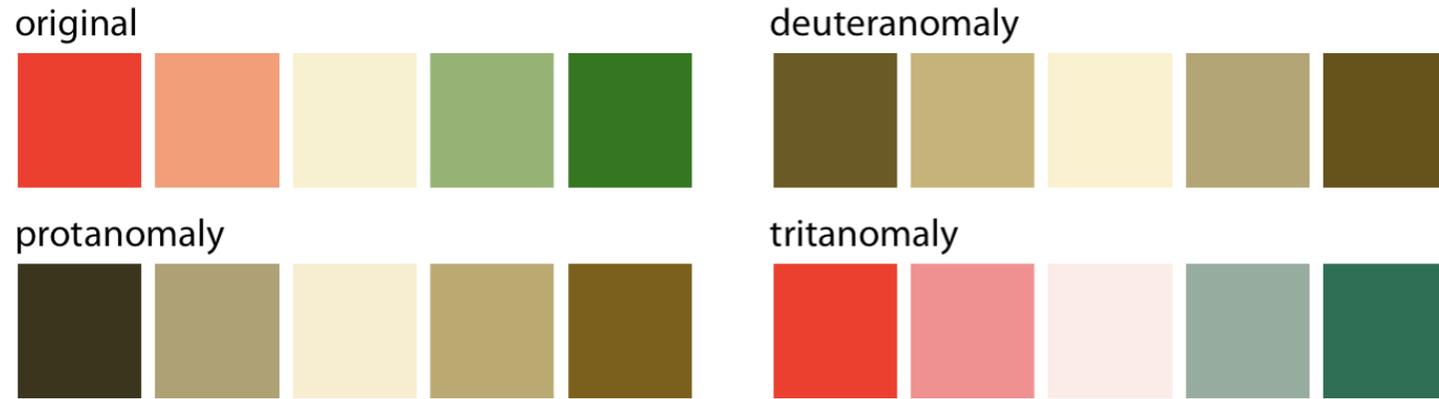
tritanomaly

Figure 19.7: A red–green contrast becomes indistinguishable under red–green cvd (deuteranomaly or protanomaly).

original

deuteranomaly

protanomaly
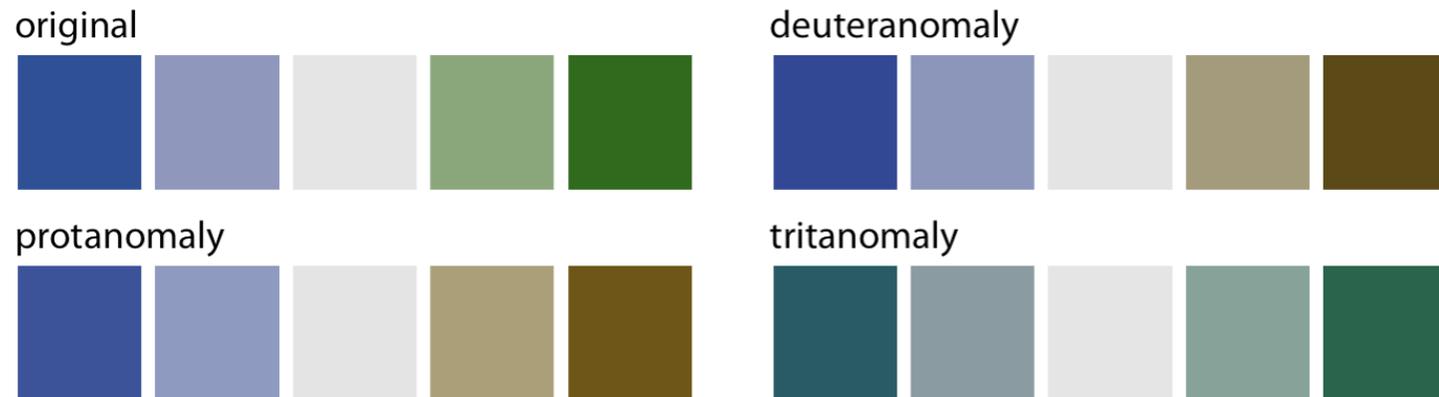
tritanomaly

Figure 19.8: A blue–green contrast becomes indistinguishable under blue–yellow cvd (tritanomaly).

https://clauswilke.com/dataviz/color-pitfalls.html#not-designing-for-color-vision-deficiency

Design for colour-vision deficiency

https://colorbrewer2.org

https://clauswilke.com/dataviz/color-pitfalls.html#not-designing-for-color-vision-deficiency

Treat your colours fairly*

*unless you deliberately don't want to

NOT IDEAL

BETTER

Treat your colours fairly*

*unless you deliberately don't want to

rainbow scale

rainbow converted to grayscale

https://colorbrewer2.org

Accessibility

↘ Consider using an extension like [colourblinding](#) to check colours

↘ Use shapes/line types in addition to colour

↘ Use alt-text and appropriate legends to explain your figures

↘ Use bigger font size for titles/labels

Accessibility

Making your data viz

The best visualization software is the one that allows you to make the figures you need.

Excel

- **Good for:**
  - Quick basic charts (bar charts, scattercharts)
  - Can make dashboards
  - Using data you already have in excel
- **Limitations:**
  - Data has to be compiled/formatted, ready for visualisation
  - Some limits to types of charts
- **Cost:** Office 365 free for charities
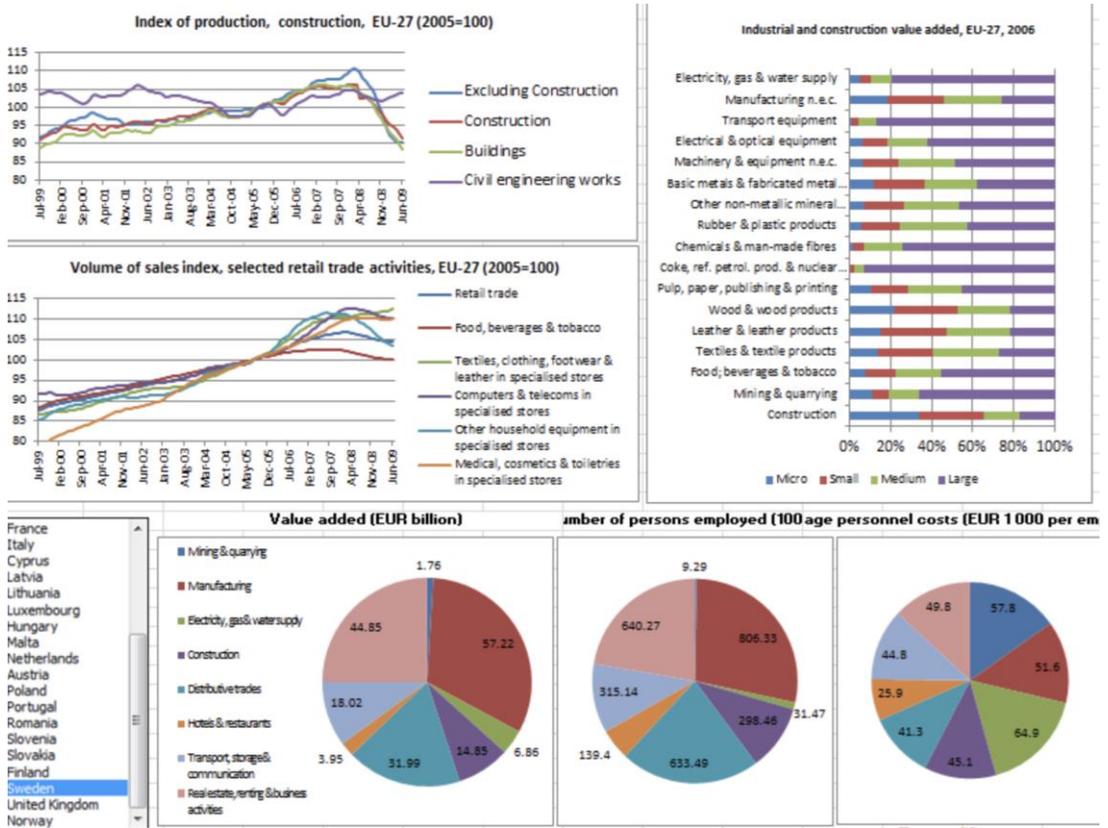
- **Good for:**
  - Dashboards
  - Maps
  - Interactivity
  - Multiple charts
  - Can be embed in websites
- **Limitations:**
  - Might be more than you need
- **Cost:** Tableau Desktop free for charities

https://public.tableau.com/en-gb/gallery/?tab=viz-of-the-day&type=viz-of-the-day

Data Studio

- **Good for:**
  - Dashboards
  - Integrates with other google tools
- **Limitations:**
  - Fewer visualisation options than Tableau

- **Cost:** Gsuite is free for charities

https://datastudio.google.com/navigation/reporting

Marketing Website Summary — Data From Google Analytics

Users	Sessions	Pageviews	Bounce Rate
49,384	61,191	305,000	51.8%
-18.7%	-24.9%	-30.1%	17.9%

How are site sessions trending?

What are the top countries by sessions?

Flourish
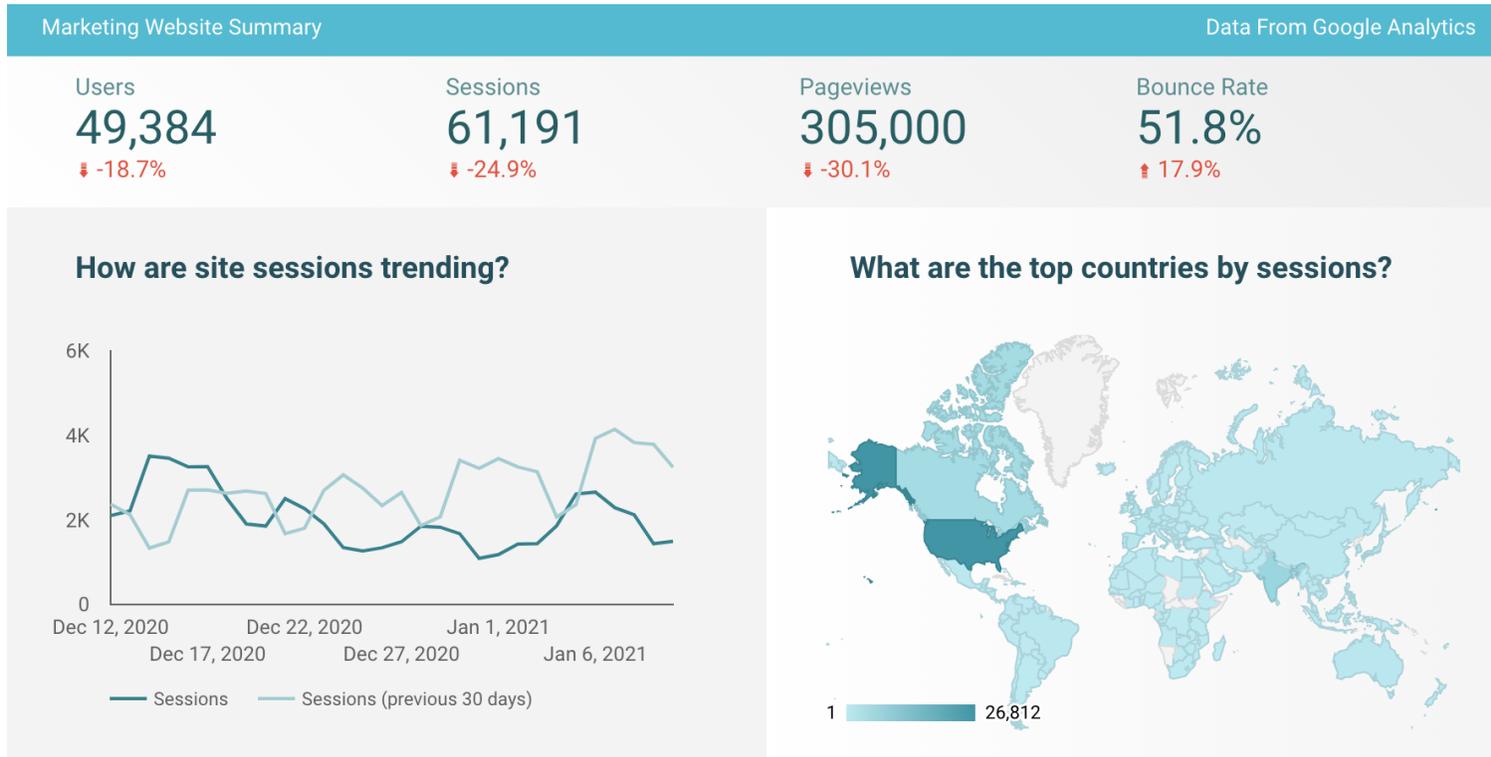
- **Good for:**
  - Storytelling and presentations
  - Interactive charts
  - Embedding in websites
- **Limitations:**
  - Free plan makes data and projects public
- **Cost:** Free option, non-profit discount

Top 15 NHS Trusts with Covid-19 staff shortages

Day-by-day cumulative numbers of staff on Covid-19 related sick leave by NHS Trust

■ East of England ■ London ■ Midlands ■ North East and Yorkshire ■ North West ■ South East ■ South West

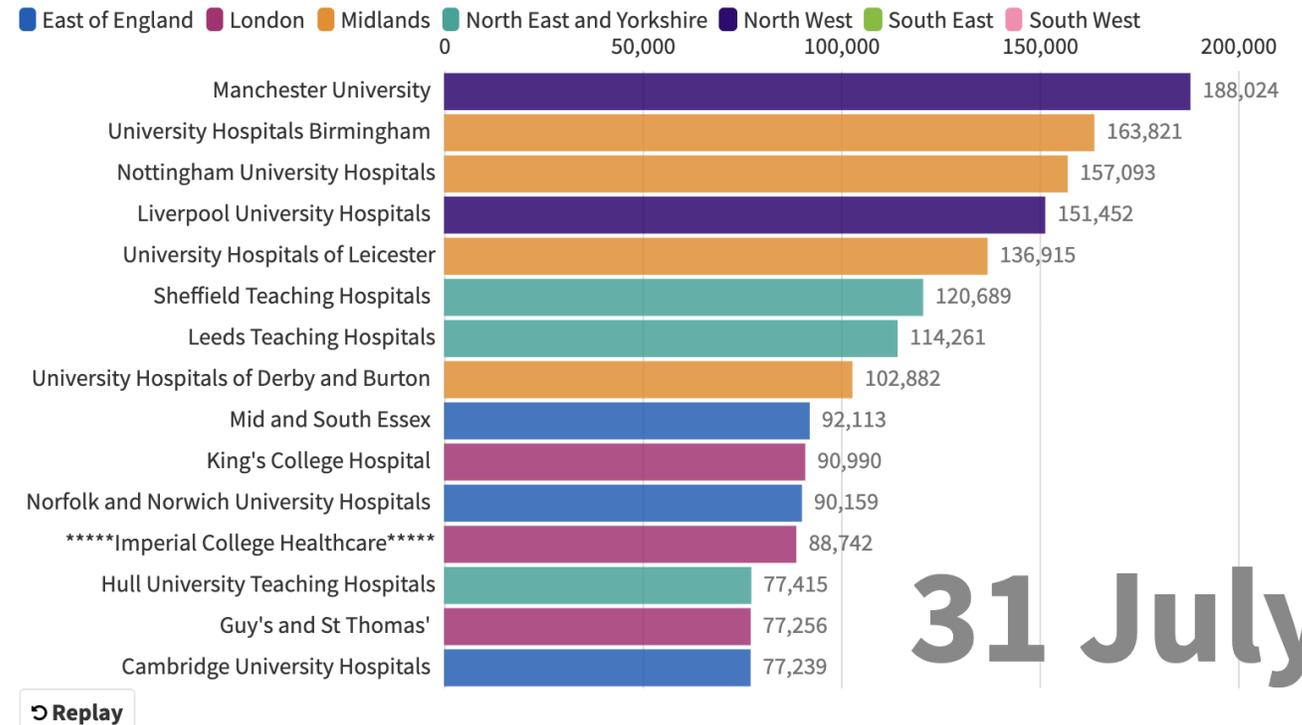| Trust	Value
Manchester University	188,024
University Hospitals Birmingham	163,821
Nottingham University Hospitals	157,093
Liverpool University Hospitals	151,452
University Hospitals of Leicester	136,915
Sheffield Teaching Hospitals	120,689
Leeds Teaching Hospitals	114,261
University Hospitals of Derby and Burton	102,882
Mid and South Essex	92,113
King's College Hospital	90,990
Norfolk and Norwich University Hospitals	90,159
*****Imperial College Healthcare*****	88,742
Hull University Teaching Hospitals	77,415
Guy's and St Thomas'	77,256
Cambridge University Hospitals	77,239

↺ Replay

31 July

A Flourish bar chart race

https://www.threesixtygiving.org/2020/11/19/my-first-data-expedition-a-workshop-without-a-post-it-note-in-sight/

- **Good for:**
  - Helps you choose your graph
  - Simple charts/infographics
  - Drag-and-drop
- **Limitations:**
  - Data pasted from spreadsheet

- **Cost:** Free for non-profits

https://www.canva.com/graphs/
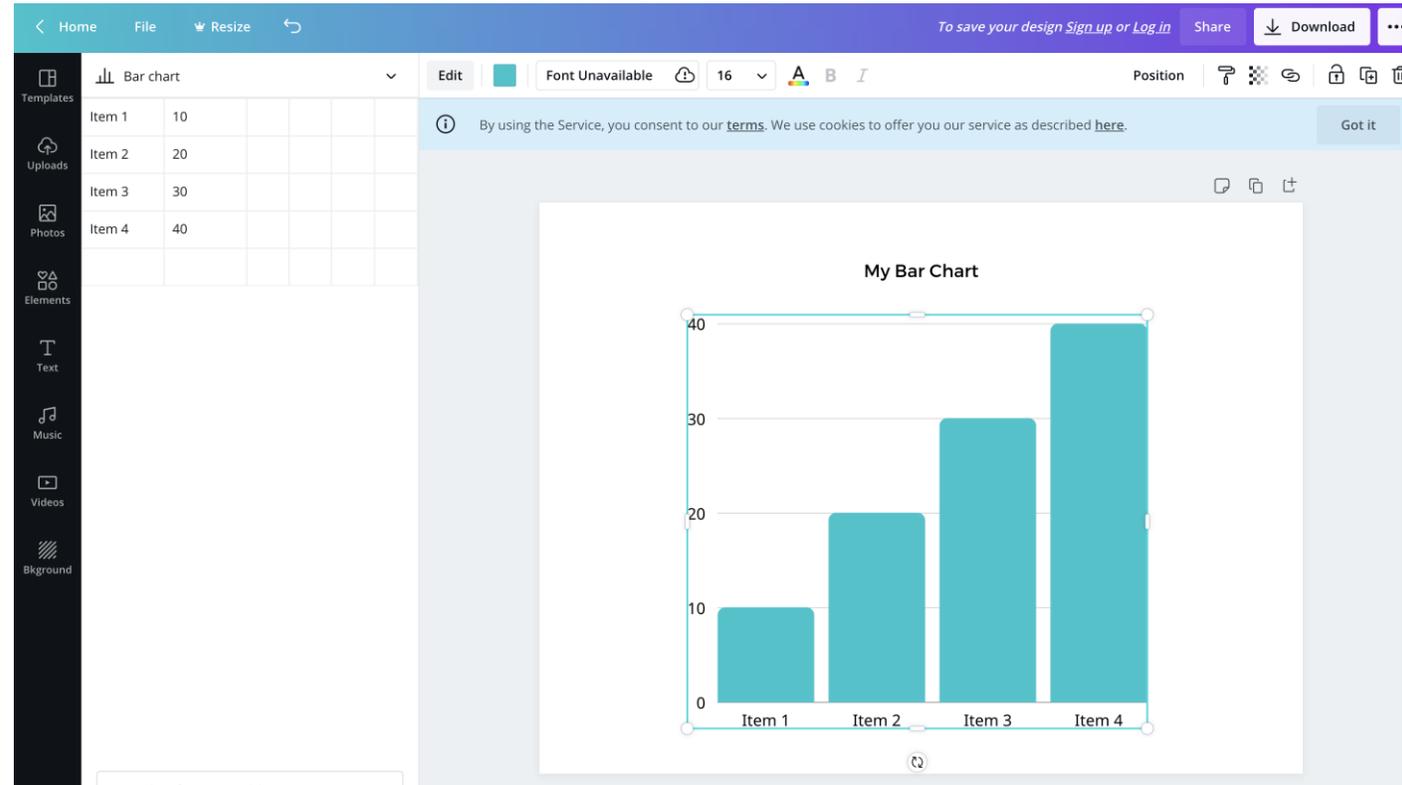
infogram

- **Good for:**
  - Interactive charts and maps
  - Presentations
  - Drag and drop
- **Limitations:**
  - Free version makes visualisations public

- **Cost:** Free option

GLOBAL WATER STRESS RANKINGS

https://infogram.com/

PIKTOCHART

- **Good for:**
  - Infographics
  - Posters
  - Drag and drop
- **Limitations:**
  - Less suited to integrating with data sources
- **Cost:** Free option, non-profit discount

Webinar

Report

Registrant to attendee conversion

*Marketing events only

👤 63%

When do they register

Register the week of live webinar
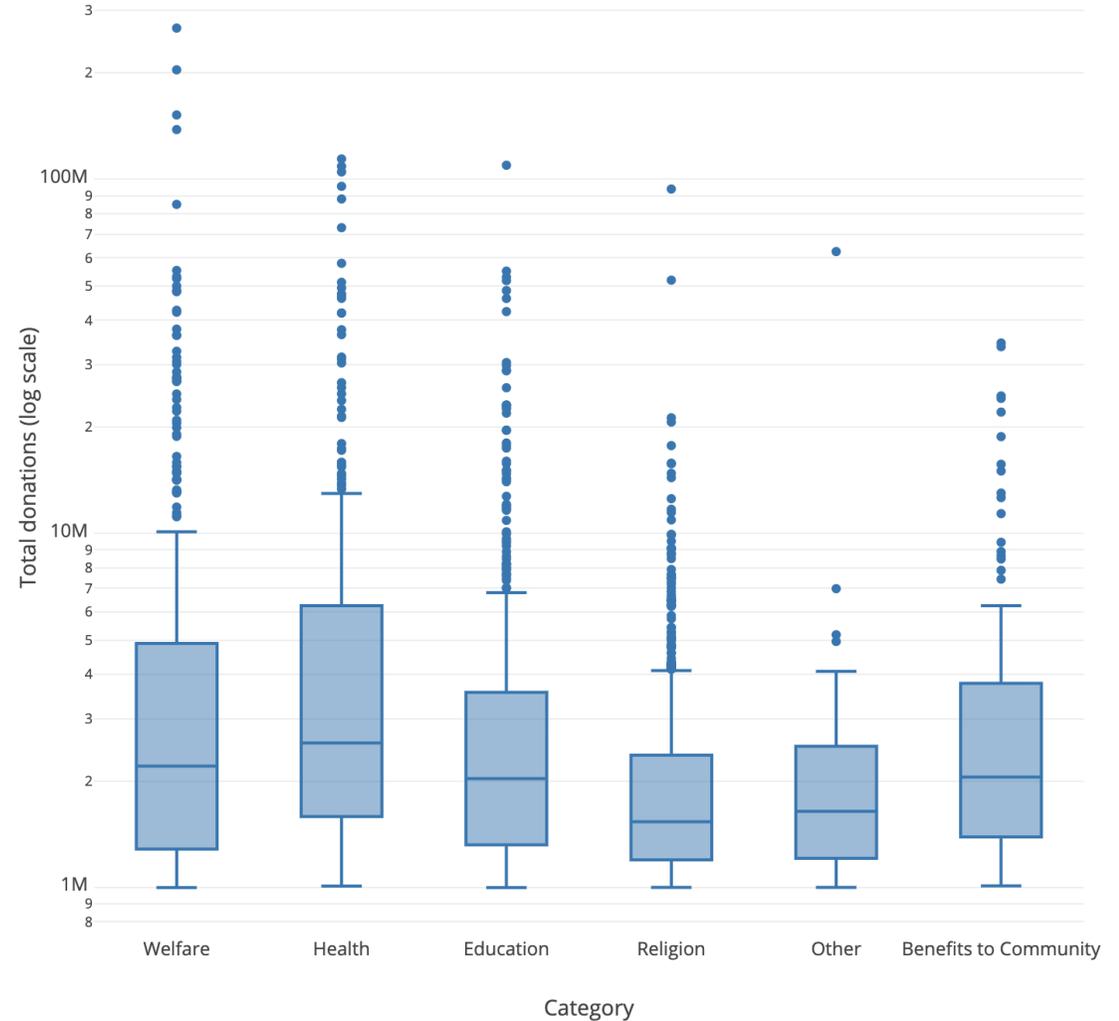
34%

https://piktochart.com/

plotly

- **Good for:**
  - Interactive charts
  - Embedding in websites
  - Attractive default formats
- **Limitations:**
  - Paid options is very expensive
- **Cost:** Free option, but limited to public charts

Canadian Charity Snapshot (donations of more than $1M)

https://plotly.com/chart-studio/

Datawrapper

- **Good for:**
  - Quick charts, maps, and tables
  - Interactive
  - Attractive default formats
- **Limitations:**
  - Limited formatting options with free version
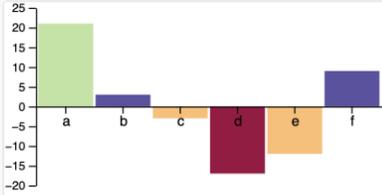- **Cost:** Free option

https://www.datawrapper.de/
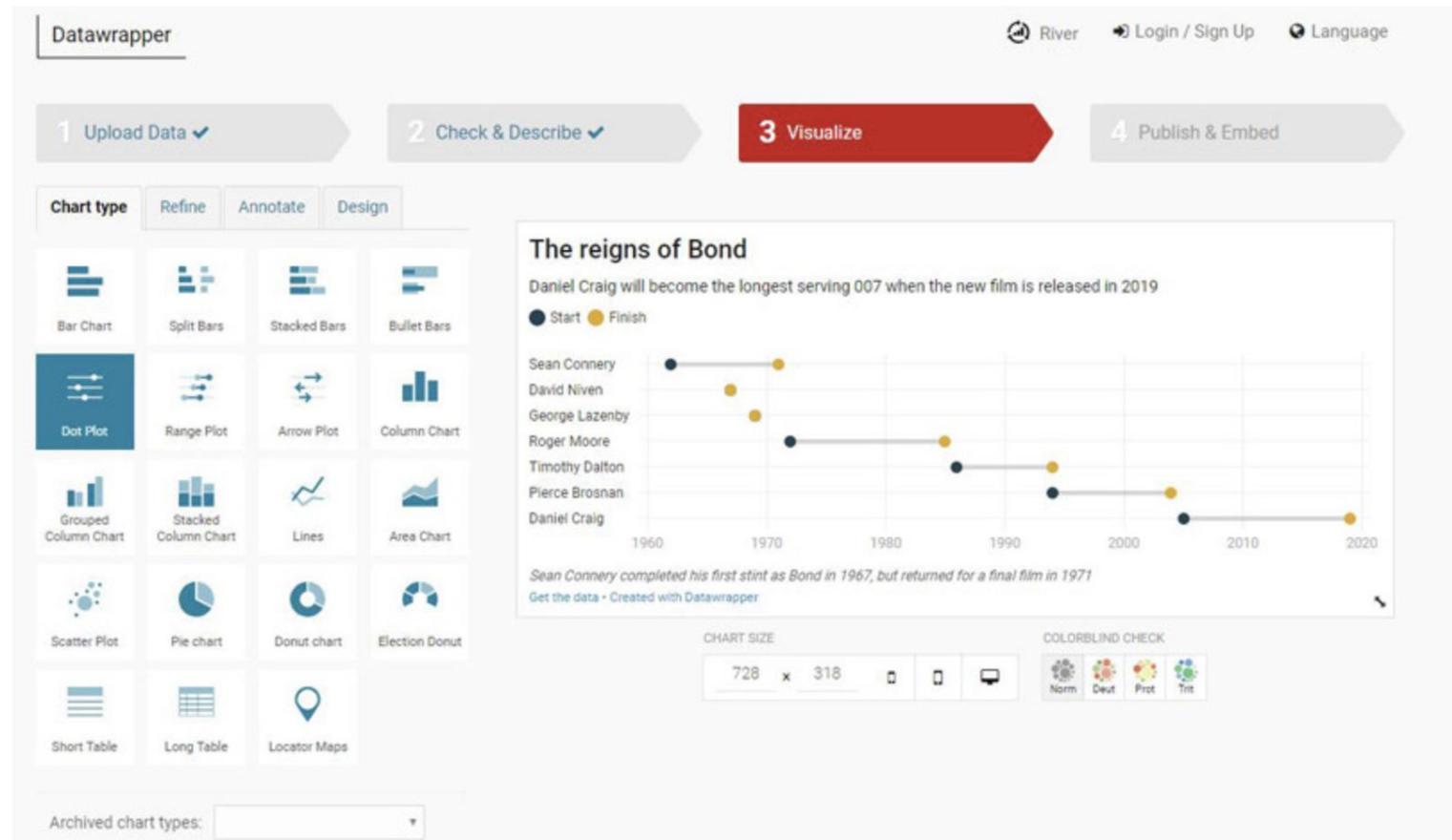
RAWGraphs

- **Good for:**
  - Lots of chart option
  - Can import data
  - Decent formatting options
- **Limitations:**
  - No maps or interac figures
- **Cost:** Free

2. Choose a chart

Show All charts ▾

Bar chart

It displays a categorical dimension and related amounts. Each bar represents a category, width is proportional to the quantitative dimension.

∞ Code   ∞ Tutorial

Alluvial Diagram Correlations, proportions	**Arc Diagram** Networks
Multi-set bar chart Correlations, proportions	**Stacked bar chart** Correlations, proportions
Box plot Distributions	**Bubble chart** Correlations, proportions
Circle Packing Hierarchies, proportions	**Circular dendrogram** Hierarchies, proportions
Convex hull Correlations, proportions	**Linear dendrogram** Hierarchies, proportions
Line chart Time series, correlations	**Matrix Plot** Correlations, time series, proportions
Sankey Diagram Flows, networks	**Streamgraph (area chart)** Correlations, proportions
Treemap Hierarchies, proportions	**Violin plot** Distributions

Bar chart Correlations

Beeswarm plot Distributions, time series, proportions

Bumpchart Correlations, proportions

Contour plot Correlation, density

Hexagonal binning Correlation, density

Radar Chart Correlations

Sunburst diagram Hierarchies, proportions

Datawrapper

- **Datawrapper:** https://www.datawrapper.de/

- **Data:**
  https://docs.google.com/spreadsheets/d/1ZUR1tIAzfU1vxwOhYrVy0Z_Bebo
  kX63NSNnPIR2MmQo/edit?usp=sharing

Resources

General guides

https://clauswilke.com/dataviz/index.html

https://www.toptal.com/designers/data-visualization/data-visualization-best-practices

https://depictdatastudio.com/data-visualization-design-process-step-by-step-guide-for-beginners/

https://towardsdatascience.com/10-free-tools-to-instantly-get-started-with-data-visualisation-d7fadb5f6dce

Choosing a chart

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

https://datavizproject.com/

Accessibility

https://medium.com/nightingale/writing-alt-text-for-data-visualization-2a218ef43f81

https://fossheim.io/writing/posts/accessible-dataviz-design/

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

https://support.infogram.com/hc/en-us/articles/360013046634-Creating-accessible-content

https://accessibility.blog.gov.uk/2016/09/02/dos-and-donts-on-designing-for-accessibility/