Data Visualization Report

According to this site, data visualization is "the practice of translating information into a visual context, to make data easier for the human brain to understand and pull insights from". Given that 65% of the world's population are visual learners, visuals can help us:

- Identify patterns
- Analyse information/draw informed conclusions
- Quantify values/influences
- Capture an audience's attention by making information understandable

As summarised here, "well presented data forms the backbone of a compelling story. It has the power to strengthen and illuminate a narrative".

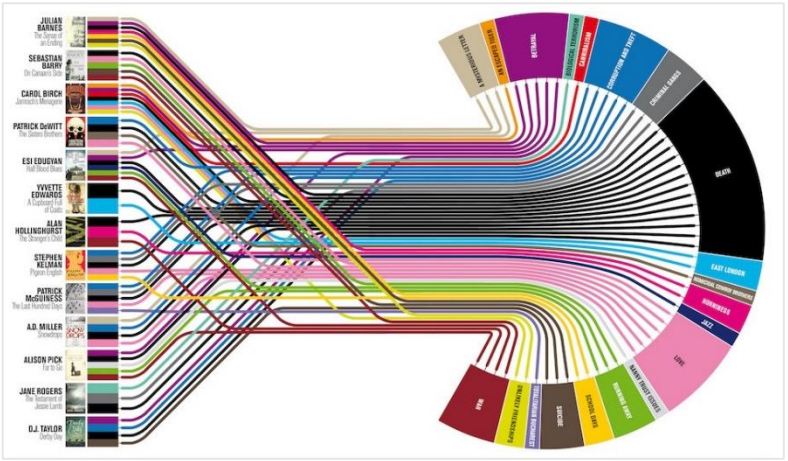
Principle 1: Simplify

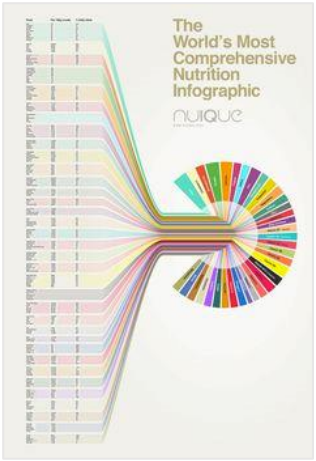
Overview

Given that the point of data visualisation is to make complex datasets easy to follow, the visuals should reflect this in their simplicity.

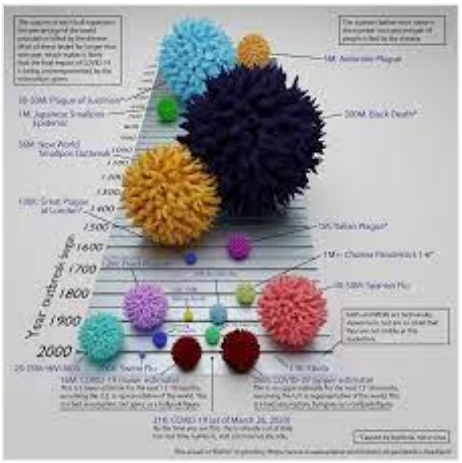
Some ways of ensuring this are:

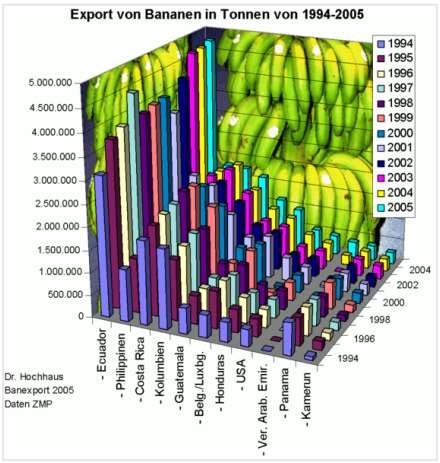
1. **Excluding extraneous information**. Only information that enhances the visual should be included, and **not too much data should be presented**. More data = more complexity.



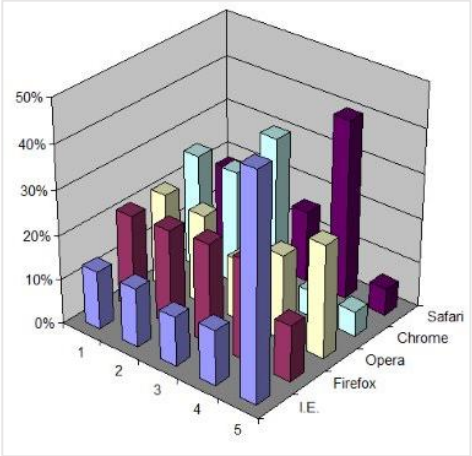


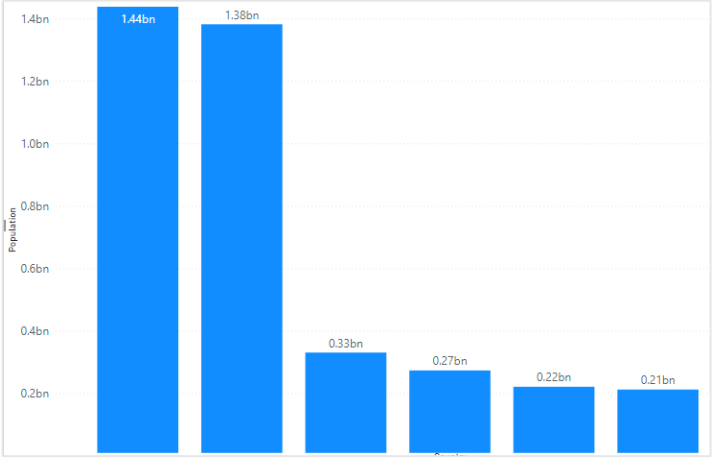
2. **Avoiding gratuitous effects** (such as 3D/shadows). If an effect is added because it looks impressive, without aiding understanding, it should be avoided.



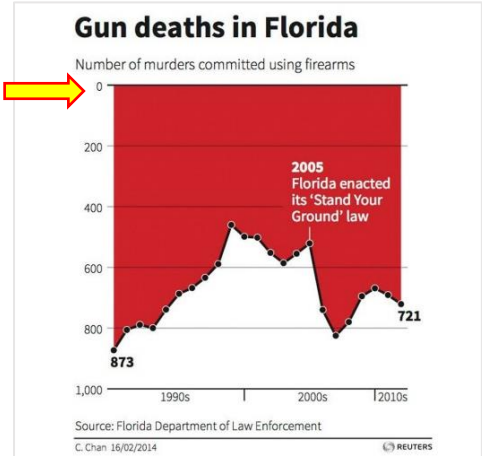


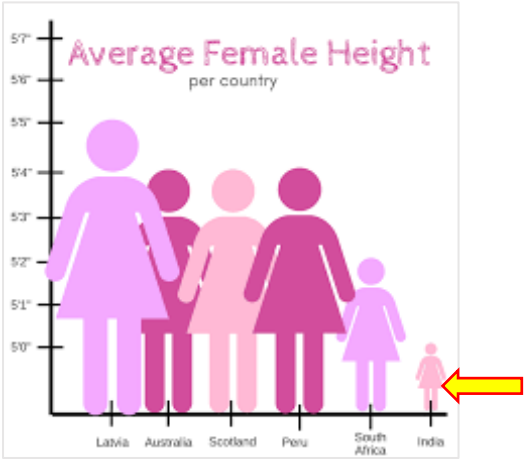
3. **Including titles/labelling axes etc.** If a user needs to check what is being presented, there is an issue. Visuals should be near 'stand-alone'.





4. **Not having misleading visuals**, via incorrect scaling or misleading axes.





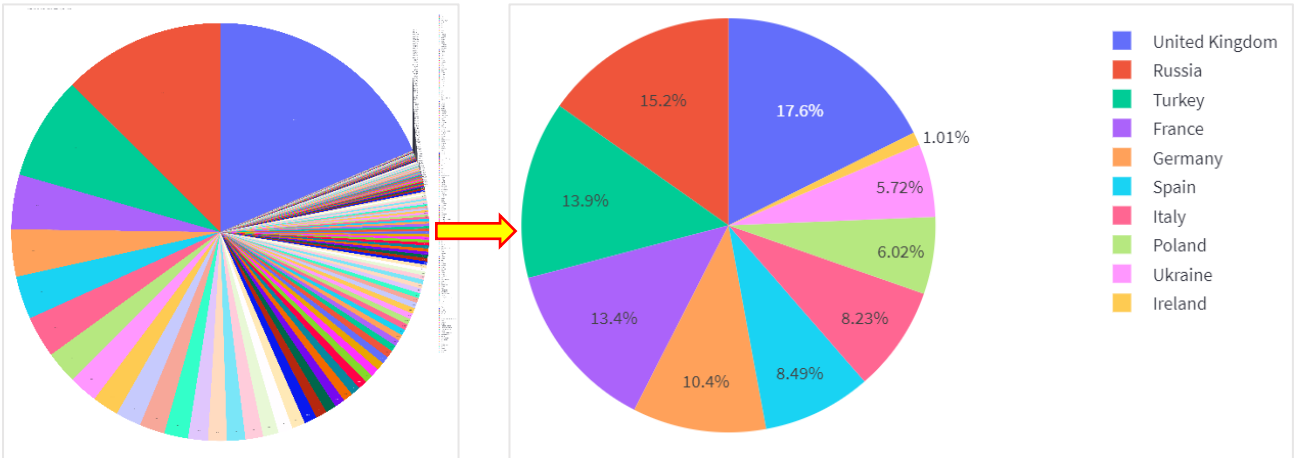
Overcomplicating visuals can result in the message being missed (e.g., admiring effects meaning the message is lost), or misinterpreted.

Compliance

My aim was to make my visuals 'stand-alone'; by looking at each graph, the viewer knows what I am plotting, without background information.

Using the same numbering system as before:

1. Began with 196 countries, split into regions, resulting in 280 data points. Plotting these together resulted in 'crowded' visuals, so I eliminated the regions and **reduced my dataset** to include 10 countries.

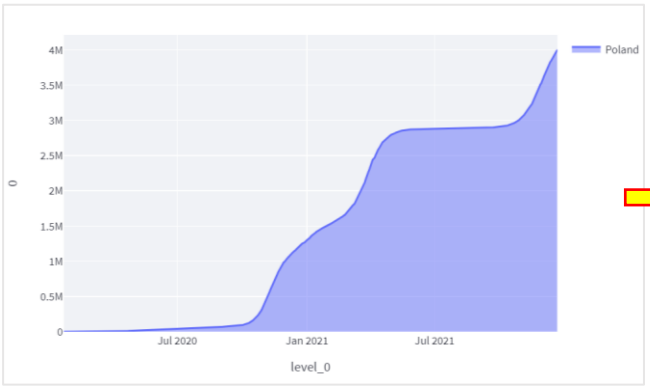


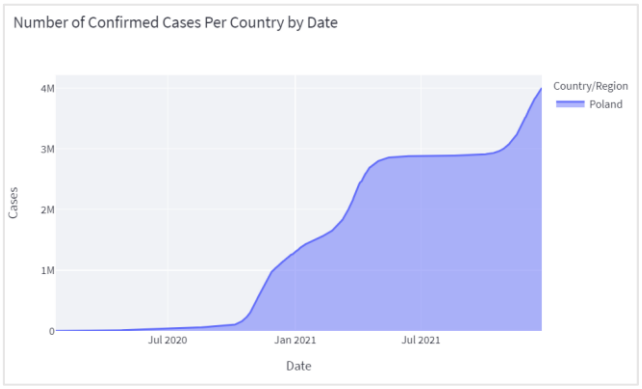
2. Ensured that each visual was on **a plain background**, with no distracting effects.

Please note that only the worst 9 European countries (as of 22/12/21 - United Kingdom, Russia, Turkey, France, Germany, Spain, Italy, Poland, Ukraine) plus Ireland are included in this analysis

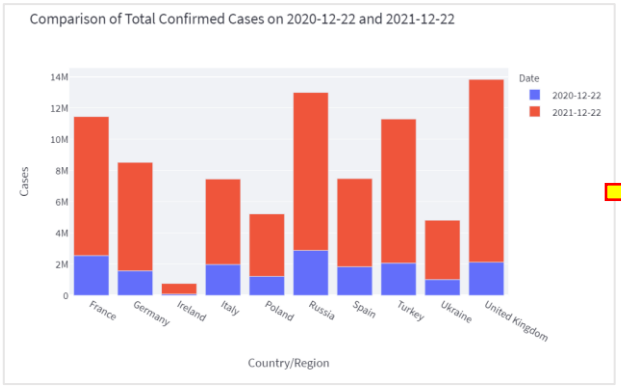
	2020-01-22	2020-01-23	2020-01-24	2020-01-25	2020-01-26	202
France	0	0	2	3	3	
Germany	0	0	0	0	0	
Ireland	0	0	0	0	0	
Italy	0	0	0	0	0	
Poland	0	0	0	0	0	
Russia	0	0	0	0	0	
Spain	0	0	0	0	0	
Turkey	0	0	0	0	0	
Ukraine	0	0	0	0	0	
United Kingdom	0	0	0	0	0	

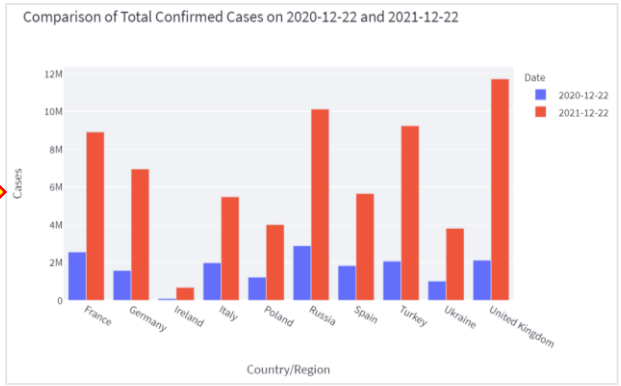
3. I ensured that **every plot had a title**. Where it was unclear, I **relabelled axes/colour guides**, so it's easy to understand what is being plotted.





4. **Avoided misleading images, shapes, sizes** that may encourage inaccurate conclusions.



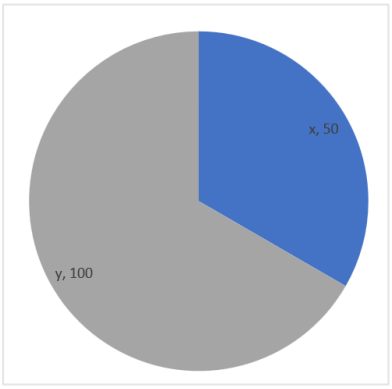


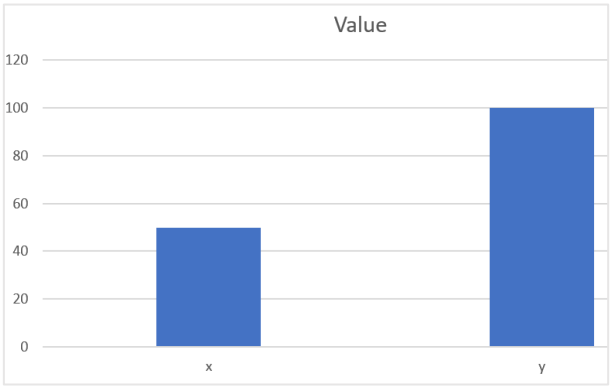
Principle 2: Understand Magnitudes

Overview

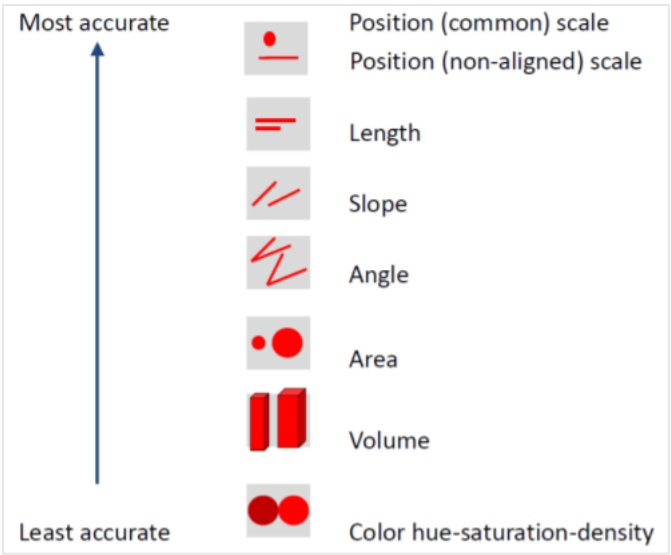
The basis for this principle is Weber's Law, combined with Stephen's Power Law. Effectively, these quantify quite how different two metrics must be, for it to be noticeable (e.g., we easily differentiate between an acute/obtuse angle, but can't differentiate between 90°/91°).

From here, different ways of presenting data are more effective in conveying relative magnitudes. For example, comparing the same 2 values (50, 100) using a pie/bar chart:





In the bar chart, we can easily see that 100 is twice 50. In the pie chart, 100 is still clearly bigger than 50, but doesn't look *double* it. This means the bar chart is easier to understand (initially), summarised below:



Compliance

I tried to use the most effective graphs to present my data (ordering them accordingly in drop-downs). Every graph is on a common scale (most accurate), and colour (least accurate) is used as described later, but not to describe magnitudes.

Plotting total cases per country:

- Bar chart is optimal; utilises length to represent numbers, with colour being used as an *additional* emphasis.
- Scatter plot is next best; uses length (height) to show number of cases. Still easy to follow, but slightly less-so.

Pie chart is the least effective option to graph them together to compare cases; Ireland's 1% slice is very small (with more data it would become negligible). I still included it, as it's the most effective way of showing total cases in countries *relative to each other*, versus on the same scale side-by-side.

Plotting cases over time:

- Line graph is most effective, as we combine length, slope and angle to show the magnitude of cases.
- Density chart would be preferable to a line graph, as then we *also* have area. However, there was no point in graphing both a line/density chart, so I created a *stacked* density chart.
    - This still has all the same advantages, but shows how each country contributes to the *total* cases (for these 10 countries), instead of each country's individual total.
    - I did not put this first as stacked graphs can be slightly confusing, if you do not realise that cases per country are being added/stacked together.

Graphing cases by date, the same principles as for graphing by country apply and I used a grouped bar chart to show the before/after by country. A scatter/pie chart would not work (comparing by date *and* country can't be - reasonably - represented).

Principle 3: Use Colour

Overview

Colour is an effective tool in data visualisation, as "using colour can increase memory, aid pattern recognition, and attract attention to priority information". Certain colours are associated with certain properties (red = danger, green = money), and can be used to create memorable visuals. Equally, colour can draw attention to important parts of our visuals (highlighting in yellow = NB).

Among other properties, colour can:

- Represent scale (e.g., height above/below sea level) using a **diverging palette**. Endpoints are very different (with neutral centre), representing a spectrum. This can be done by going from warm to cool/bright to dark, giving a sense of two extremes, with large contrast between.

- Exemplify a pattern/continuous data (e.g., temperature of an ice cube melting) using a **sequential palette**, which "use a single colour in a variety of saturations/gradient". This gives a sense of uniformity, order, and a 'passage of time' effect.

- Represent categories (e.g., country) using a **categorical palette**, where colours are very distinct from each other, making them easily distinguishable. This makes them easy to both tell apart and compare them to the key.

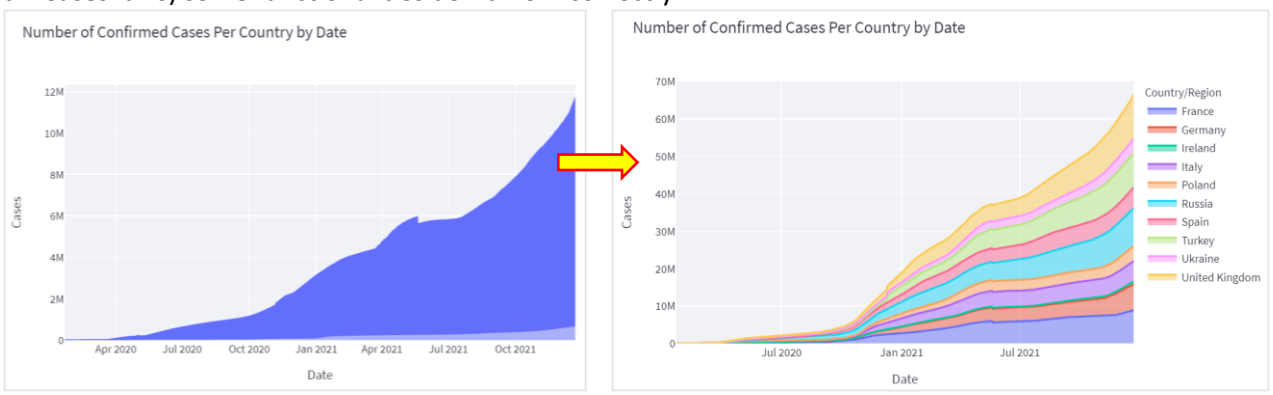
Compliance

I used the above guidelines throughout my graphs, catering to what I was trying to emphasise in each visual. I ended up having 3 needs for colour: case numbers, country, date.

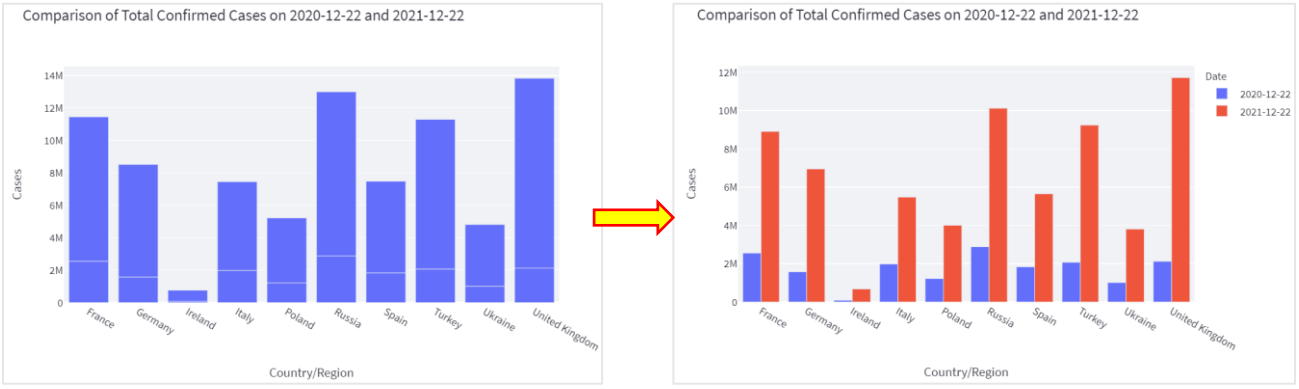
- I used divergent palettes to show scale between case numbers. Without colour, only bar height shows this, and the colours *emphasise* it.

- While using colour to represent countries, I used a categorical palette. Without colour, it is unclear what each dot represents.

- Other times, it is not even clear that multiple countries are in the plot without colour, and (as apparent here with 'Cases' axis) some functionalities don't work correctly.



- While comparing 2 dates, I used 2 contrasting colours so that they are clearly different.

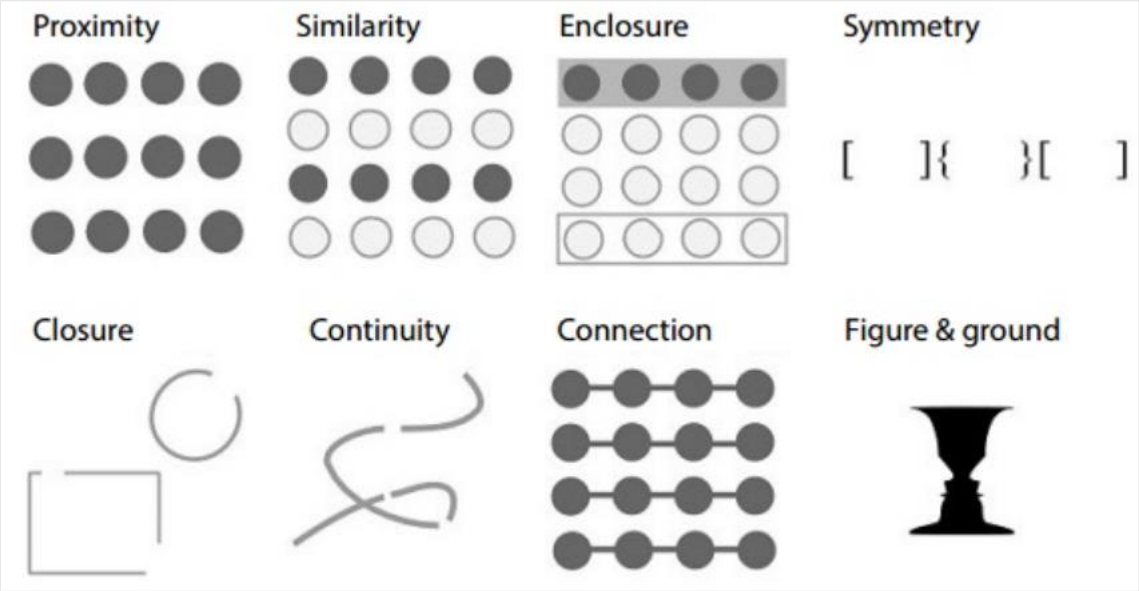


Maintaining simplicity: before adding colour, I was mindful of whether this enhances my plots. In most graphs, I had 3 pieces of information (country, cases, date), so it was actually necessary to use colour to be able to show this 3rd piece of information. I also always labelled what each colour represented.

Principle 4: Use Structure

Overview

Based on the Gestalt Psychology principles, effectively interpreting how we are programmed to interpret visuals, based on various criteria:



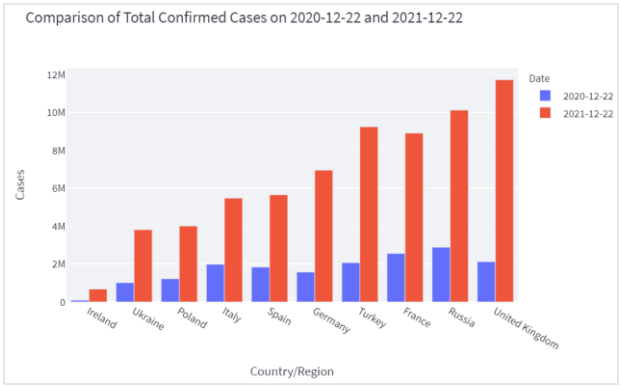
- **Proximity**: horizontal/vertical spacing affects how we interpret images.
  - Here, we see 3 rows (not 4 columns) of dots.
- **Similarity**: upon noticing patterns/similarities/trends between objects, we group them together.
  - Here, we group by colour.
- **Enclosure**: if objects are in boxes, we see them as being separate.
  - Here, the first/last rows are interpreted as separate.
- **Symmetry**: instead of seeing individual units, if a 'matching pair' exists, we group them together as one object.
  - Here, we don't see 6 individual brackets, we see 3 pairs.
- **Closure**: disconnections in paths are automatically 'filled in' to create whole units.
  - Here, we don't see 3 separate, disconnected paths, we see a square and a circle.
- **Continuity**: similarly, a path will automatically be 'joined up' so it is continuous, without breaks.
  - Here, we don't see 3 separate paths, we see 1 continuous one.
- **Connection**: if different objects are connected, we consider them part of the same group.
  - Here, connected dots are considered part of the same group.
- **Figure & ground**: "*We either notice the two faces, or the vase. Whichever we notice becomes the figure, and the other the ground*".

Compliance

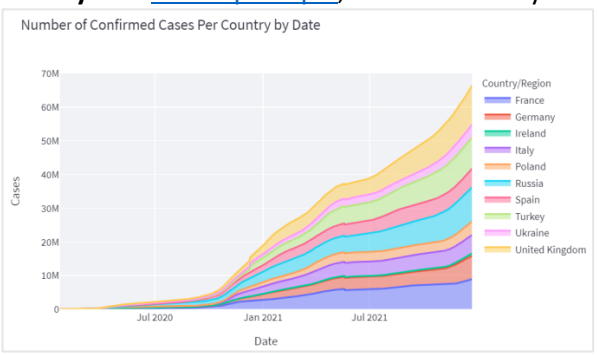
I could not find any examples of closure, continuity or 'figure and ground' in my visuals.

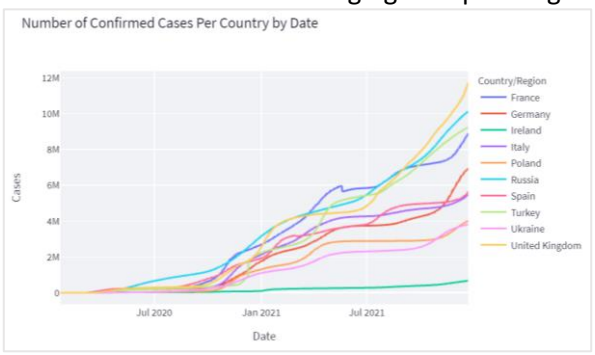
Some ways that I implemented the remaining concepts are:

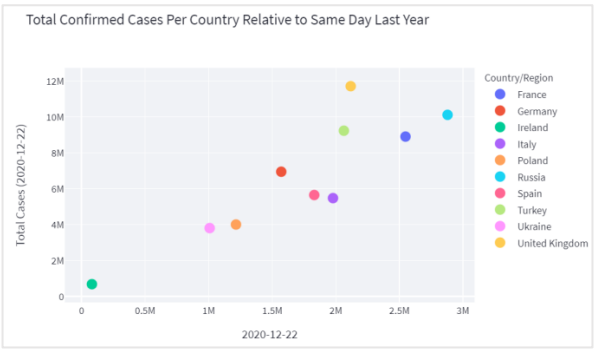
- **Proximity:** the grouped (red/blue) bars being closer together makes it clear that they're related. We automatically know 'closer to the top = higher cases'.

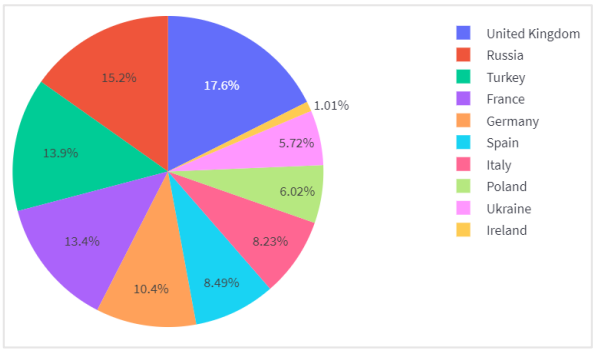


- **Similarity:** like colour principle, we immediately recognise different colours as belonging to separate groups.







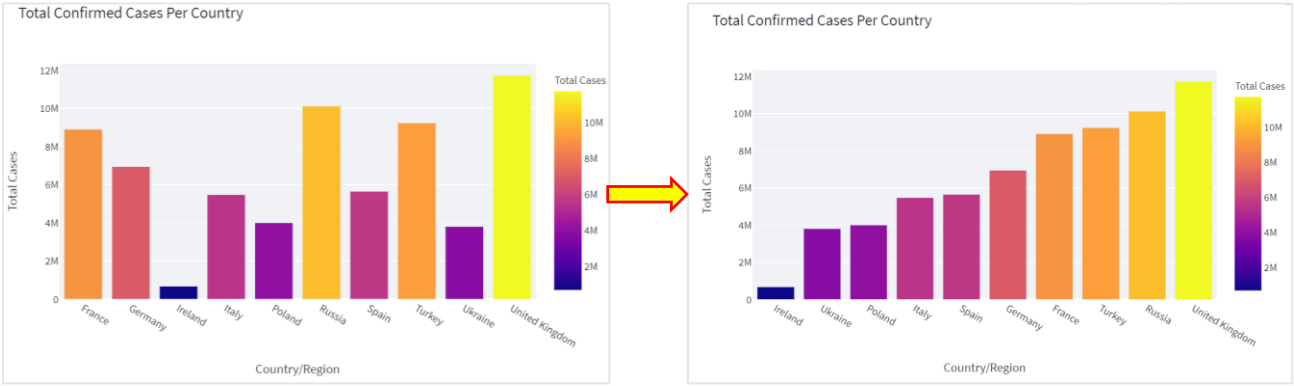


- **Enclosure:** these tables are in boxes, marking them as a different 'section' enclosed within.
  - Same with images in this report.

	2020-01-22	2020-01-23	2020-01-24
France	0	0	2
Germany	0	0	0
Ireland	0	0	0
Italy	0	0	0
Poland	0	0	0
Russia	0	0	0
Spain	0	0	0
Turkey	0	0	0
Ukraine	0	0	0
United Kingdom	0	0	0

	2020-01-22	2020-01-23	2020-01-24
count	10.0000	10.0000	10.0000
mean	0.0000	0.0000	0.2000
std	0.0000	0.0000	0.6325
min	0.0000	0.0000	0.0000
25%	0.0000	0.0000	0.0000
50%	0.0000	0.0000	0.0000
75%	0.0000	0.0000	0.0000
max	0.0000	0.0000	2.0000

- **Symmetry**: this is not making my graph symmetric (doesn't make sense to), however in a similar vein I re-ordered my charts to emphasise the upward trend (as opposed unordered).
    - Even while presenting visuals in this report, it is symmetric (both in spacing/sizing).
    - This was not a principle I remembered to apply until reaching this section. I considered redoing my previous images, however as this report is to show the process of optimisation, I chose to leave them as-was.



- **Connection**: not strictly relevant to mine (adding points makes it cluttered - below). If I'd had fewer data points (like [this example] below), we'd group dots together.