RESERVE BANK OF NEW ZEALAND
TE PŪTEA MATUA

Thinking About Data Visualisation

Christopher Ball
Christopher.Ball@rbnz.govt.nz
Reserve Bank of New Zealand
2021 Economics Department Seminar Series

Outline

1. Introduction
2. Fundamentals
3. Design philosophy
4. Special charts
5. Animation & interactivity
6. Colour & perception
7. Storytelling with Data – Practical Examples

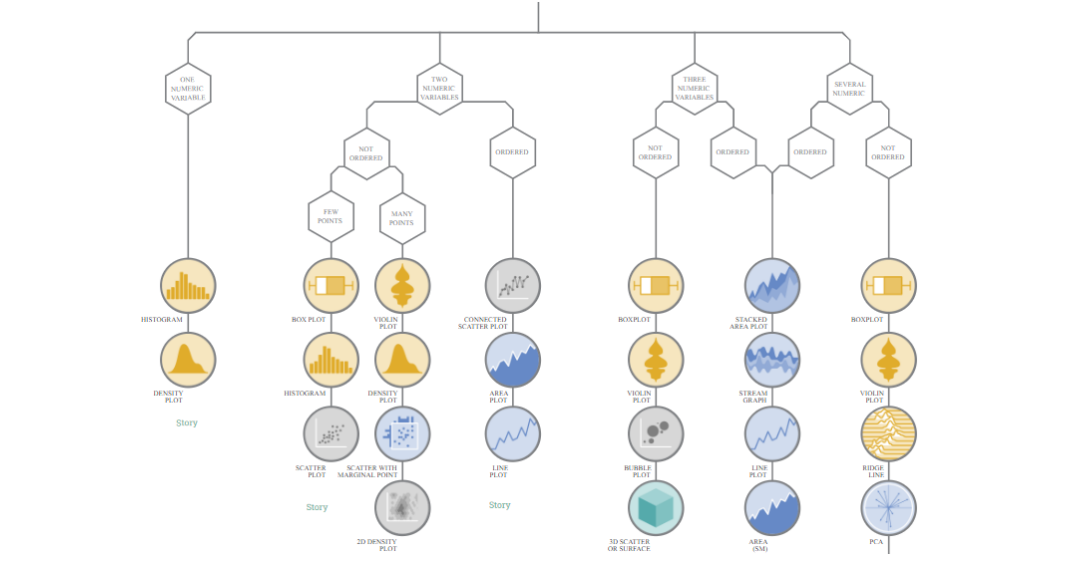
Introduction

Fundamentals

Design philosophy

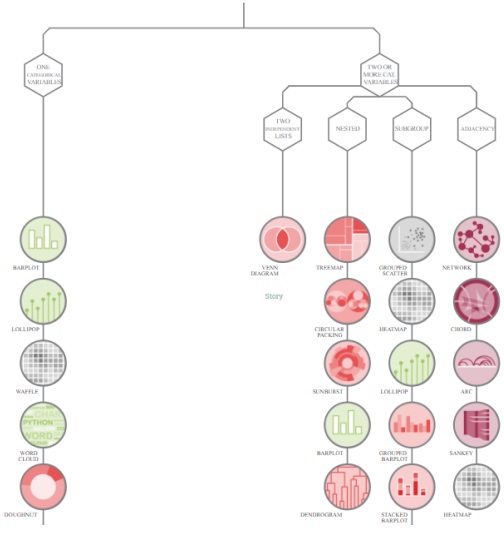
Special charts

Numeric charts



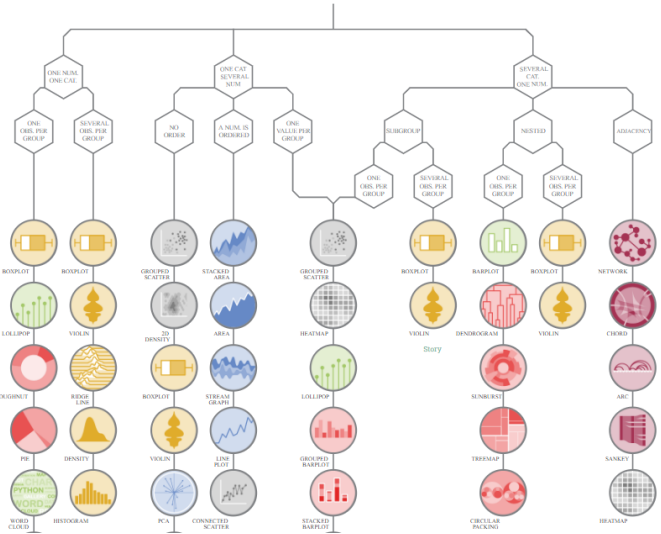
Source: https://www.data-to-viz.com/

Categoric charts



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

Numeric and categoric charts

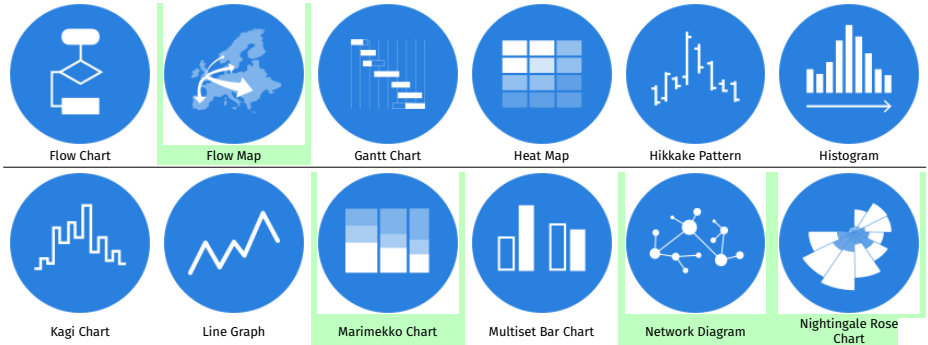


Source: https://www.data-to-viz.com/

Charts I

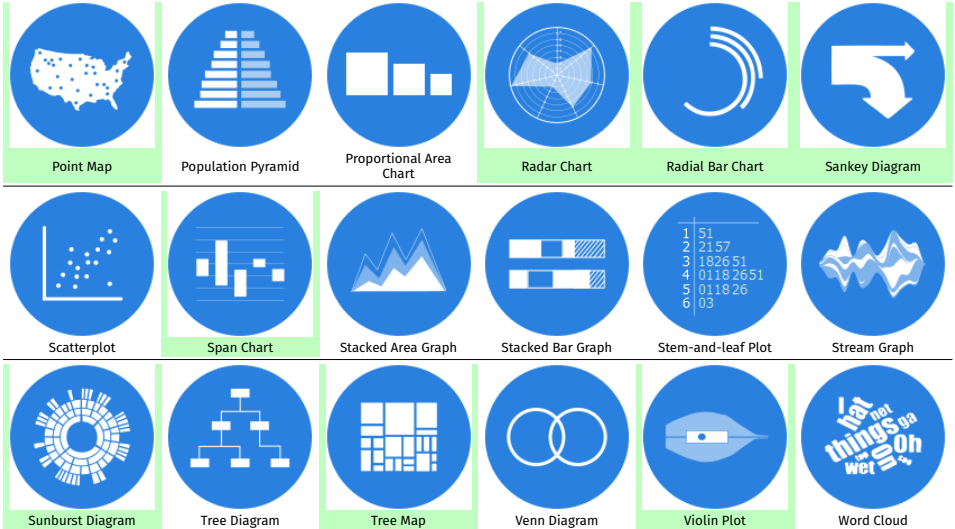
Source: https://datavizcatalogue.com/

Charts II



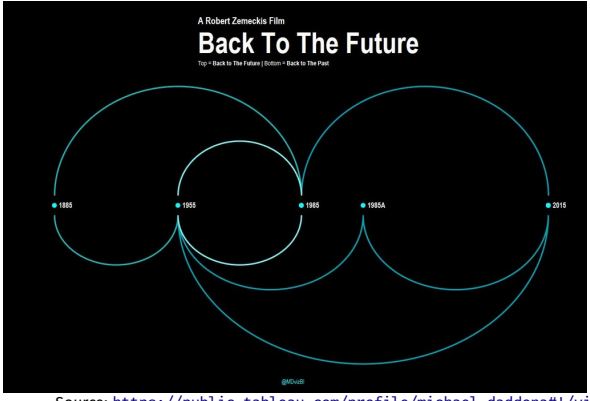
Source: https://datavizcatalogue.com/

Charts III



Source: https://datavizcatalogue.com/

Arc Diagram

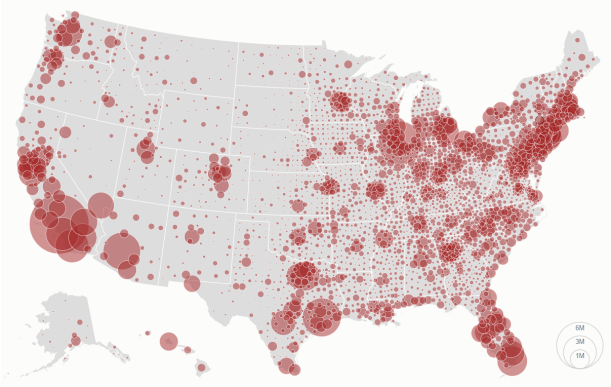


Description: An arc diagram is a special kind of network graph. Arc Diagrams are an alternate way of representing two-dimensional Network Diagrams. In Arc Diagrams, nodes are placed along a single line (a one-dimensional axis) and arcs are used to show connections between those nodes.
Data type: Relational, although typically it work better if nodes have an ordinal relationship.
Positives: Easy to display the node labels.
Negatives: They don't show structure and connections between nodes as well as 2D charts do and too many links can make the diagram hard to read due to clutter.
Comments: Sometimes known as a linear embedding. It is also NP-hard to minimise the number of crossings.

Source: https://public.tableau.com/profile/michael.daddona#!/vizhome/BackToTheFutureArcDiagram/BackToTheFutureArcDiagram

Bubble Map

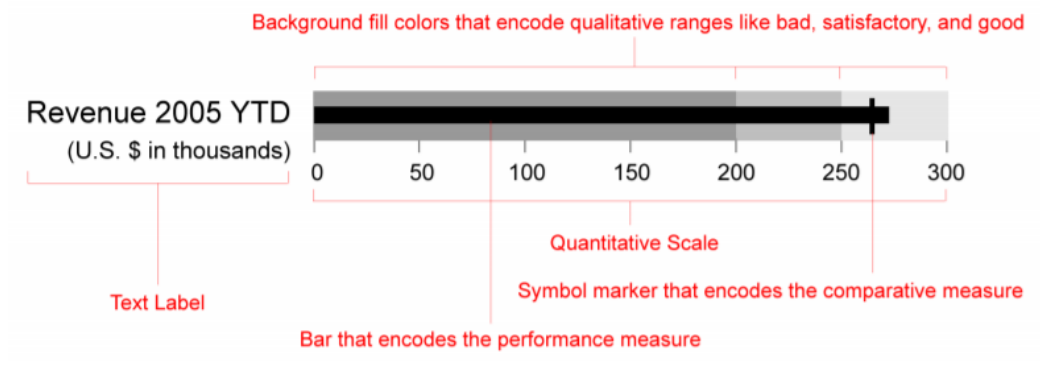
Bubble map of population by U.S. county



Source: https://bost.ocks.org/mike/bubble-map/

Description: A Bubble Map Chart is simply a combination of a bubble chart data visualization and a map. It is used to visualize location and proportaion in a simple way.
Data type: A list of geographic coordinates (longitude and latitude) and a numeric variable controling the size of the bubble.
Positives: Easy way to show geospatial trends.
Negatives: Overplotting can be an issue, even with transparency (alpha) for the bubble fill. There also needs to be sufficient discrimination in bubble sizes to be easily compared.
Comments: Need to use area rather than radius for accurate visual scaling. Interactivity is useful for bubble maps. It can be used for zooming on a specific region, or click bubbles for more information.

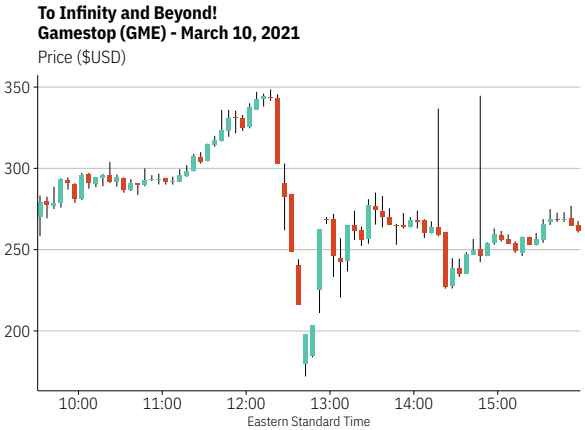
Bullet Chart



A standard bullet graph with each of its parts labeled.

Source: http://www.perceptualedge.com/articles/misc/Bullet_Graph_Design_Spec.pdf

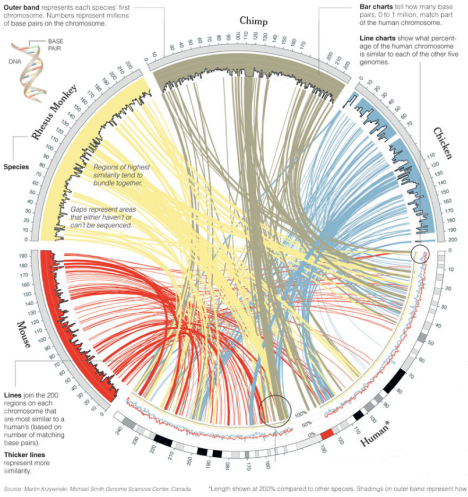
Candlestick Chart



Source: https://www.alphavantage.co/

Description: Typically used to show financial time series data, with the range for a given time period shown by a line, the opening and closing values shown in a rectangle and the colour of the rectangle switching based on positive and negative change.
Data type: Time series continuous *x* split into discrete intervals, continuous *y* values.
Positives: Commonly used and understood representation of financial time series, summarises range as well as open and closing values.
Negatives: Can be difficult to see small movements or higher frequency data.
Comments:

Chord Diagram

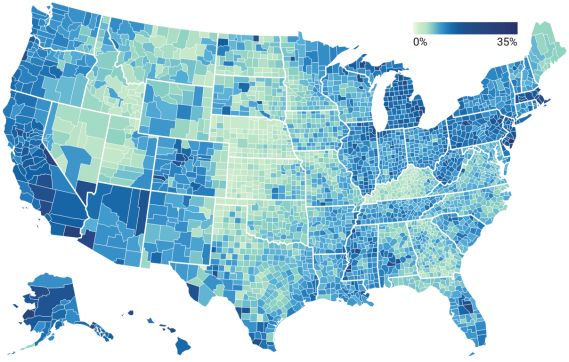


Description: Network/relation data arranged around the edge of a circle, with relationships shown as arcs (or edges).
Data type: Relational/network data
Positives: Easy-to-interpret way to display network information, commonly used in media visualisations.
Negatives: Rather difficult to automate, usually requires work on manual positioning.
Comments: Works better with large groups with links between subgroups. Can also add extra information on the edge of the circle. Non-ribbon chord diagrams essentially use the same size for each connection.

Source: https://archive.nytimes.com/www.nytimes.com/imagepages/2007/01/22/science/20070123_SCI_ILLO.html

Choropleth Map

U.S. county unemployment in June 2020



US county unemployment in June 2020

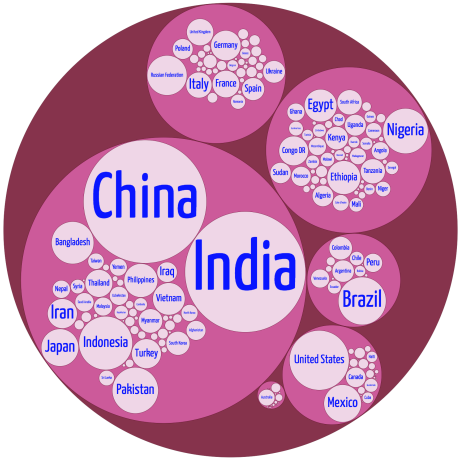
Source: BLS

Description: Geospatial data shown on a map with sub-regions coloured according to another variable.
Data type: Geospatial, typically with ordinal or continuous values by sub-region.
Positives: High data-to-ink ratio, visually intuitive way to show geosptial trends.
Negatives: Can be difficult to get the colour scale to work with larger ranges or smaller regions.
Comments: This type of chart works well with interactive charts - mouse-over can show value, location name and additional contextual information.

Source: https://academy.datawrapper.de/article/117-color-palette-for-your-map

Circle Packing

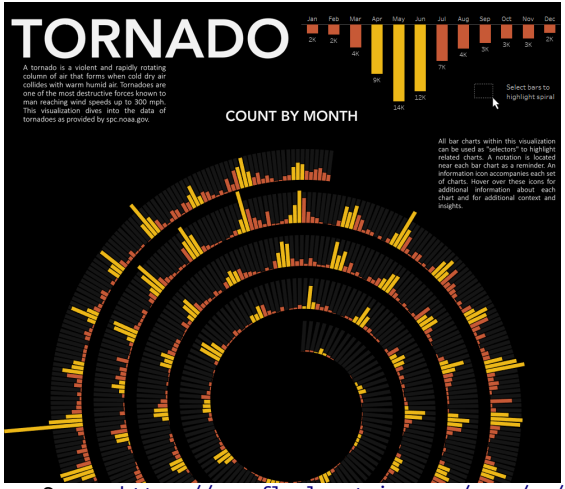
Country and continent populations (2016)



Description: Hierarchical data displayed in circles within circles (within circles…).
Data type: Hierarchical data
Positives: Not many, if any. May be more useful for Infographics instead of Data Visualisations.
Negatives: A lot of wasted space as circles have a low packing efficiency.
Comments: Alternatives like a Tree Map may be more appropriate. Algorithms for plotting are involved, so unlike most chart types it is difficult to code from first principles.

Source: https://static.packt-cdn.com/products/9781838645571/graphics/a3de031c-078e-4186-8dd1-d23bc8ba548c.png

Condegram Spiral Plot

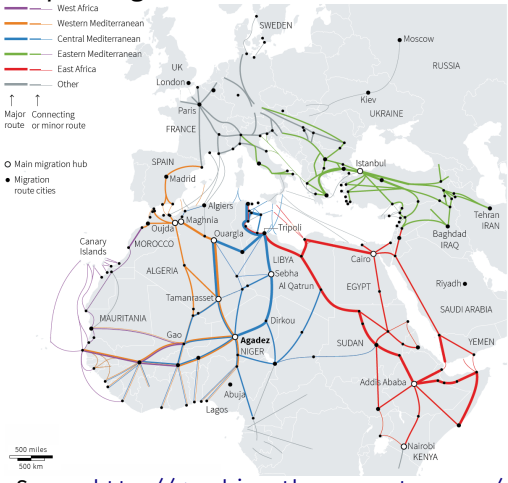


Description: Essentially a column chart wrapped into a spiral.
Data type: Time series with continuous *y* variable.
Positives: High data-to-ink ratio.
Negatives: Difficult to see seasonal patterns. *y*-axis can be difficult to read as the chart rotates.
Comments: Alternatives such as facets may be more useful.

Source: https://www.flerlagetwins.com/2020/03/how-i-created-this-spiral-chartand-why.html

Connection Map

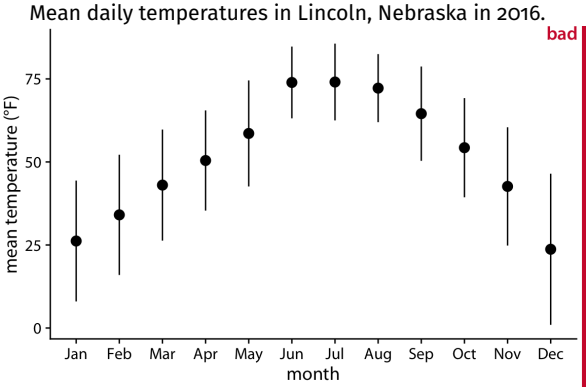
Europe's migration crisis



Description: Geospatial map with line connections drawn between places.
Data type: Geospatial data with pairs of connections.
Positives: Leading way to show geospatial connections.
Negatives: Can be challenging to interpret with a large number of lines.
Comments: A more common example is Google Maps.

Source: http://graphics.thomsonreuters.com/15/migrants/index.html

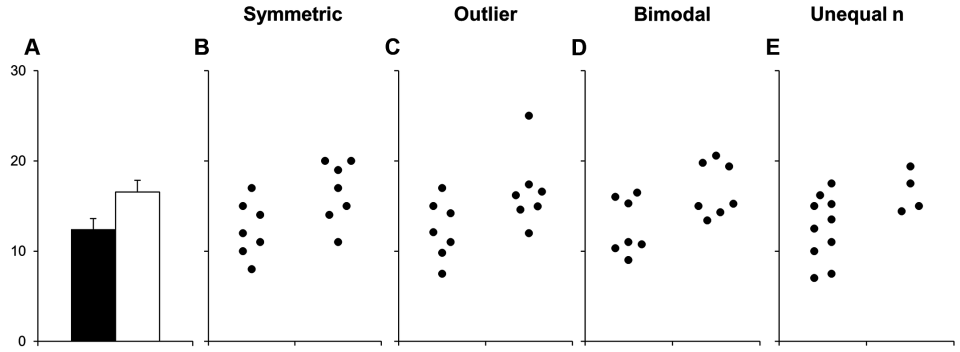
Error Bars



Description: Error bars add an uncertainty dimension to an existing chart (e.g. point, line, bar,…)
Data type: Additional dimensions of uncertainty for an existing chart (e.g. standard deviation/standard error)
Positives: Visually show an indication of uncertainty. Can be used with many existing plot types.
Negatives: Can hide the underlying distribution by simplifying to an error bar (see next slide).
Comments: Paraphrased from the source: "*First, by representing each distribution by only one point and two error bars, we are losing a lot of information about the data…it is definitely not obvious what the error bars represent. There is no commonly accepted standard.*" In the context of the above, the error bars could show the standard error (estimate of the mean) or the standard deviation (variability in the underlying data).

Source: https://clauswilke.com/dataviz/boxplots-violins.html

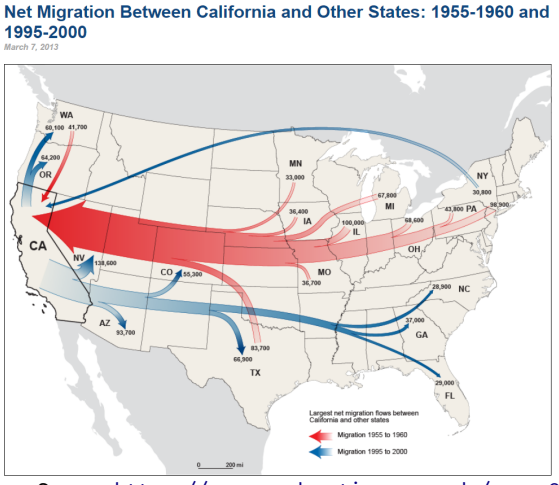
Error bar - distortion



Test	p value			
T-test: Equal var.	0.035	0.050	0.026	0.063
T-test: Unequal var.	0.035	0.050	0.026	0.035
Wilcoxon	0.054	0.073	0.128	0.103

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

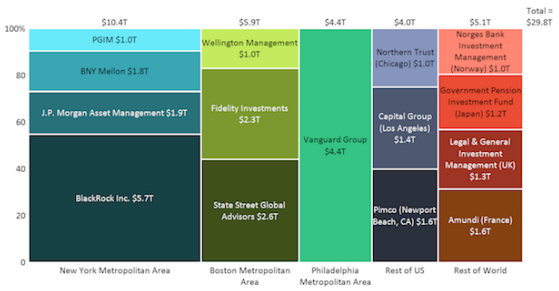
Flow Map



Description: A flow map is a map that visualizes movement between places – often across large regions, even the entire globe. Flow maps can be classified into two main types: those that represent origins and destinations, and those that map routes.
Data type: Geospatial, with pairs of coordinates or routes with edge weighting.
Positives:
Negatives: Plotting too many routes or paths can make the chart difficult to read.
Comments: Possibly the most famous flow map ever designed was drawn by Charles Minard; it represents the French army's travel and suffering during the Russian campaign of 1812. Edward Tufte described this work as perhaps the best statistical graphic that had ever been created (Next Slide – Tufte VDQI 2001).

Source: https://www.e-education.psu.edu/geog486/node/679

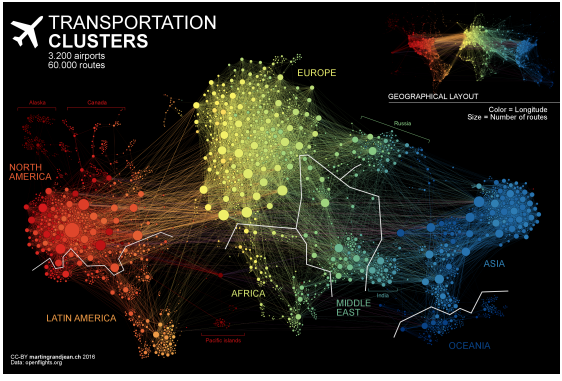
Marimekko Chart



Description: Marimekko Charts are used to visualise categorical data over a pair of variables.
Data type: In a Marimekko Chart, both axes are variable with a percentage scale, that determines both the width and height of each segment. Marimekko Charts work as a two-way 100% Stacked Bar Graph. This makes it possible to detect relationships between categories and their subcategories via the two axes.
Positives: Suited for giving a general overview of the data.
Negatives: The main flaws of Marimekko Charts are that they can be hard to read, especially when there are many segments. Also, it's hard to accurately make comparisons between each segment, as they are not all arranged next to each other along a common baseline.
Comments:

Source: https://www.indianadscompany.com/data-visualization-101-how-to-choose-the-right-chart-or-graph-for-your-data/#7_Mekko_Chart

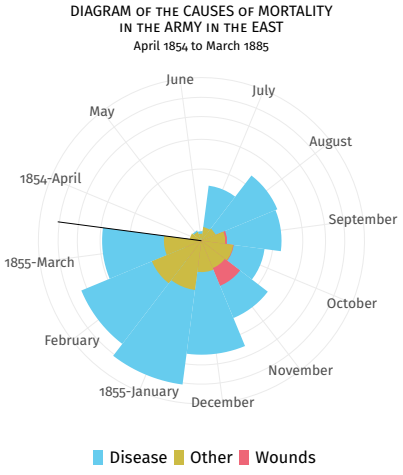
Network Diagram



Description: Network diagrams show connections between a set of entities. Each entity is represented by a Node (or vertex). Connections between nodes are represented through links (or edges).
Data type: Relational.
Positives: Can be an intuitive way to see connectedness, clusters and key nodes.
Negatives: Figures can get cluttered or unreadable if there are too many nodes or links. Data pre-processing and chart automation can also challenging.
Comments: There is enough material for an extra session discussing the intricacies of network diagrams: weighted/unweighted, directed/undirected, node layout, centrality, grouping, interactivity, etc.

Source: http://www.martingrandjean.ch/wp-content/uploads/2016/05/airports-world-network.png

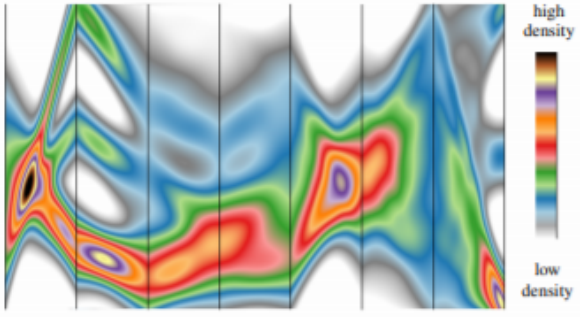
Nightingale Rose Chart



Description: Essentially this is a combination of a stacked bar/column chart to show a continuous *y* variable by group, along a continuous *x* dimension (time).
Data type: Continuous *x* (time) and *y* (death rate) by discrete groups (cause of death).
Positives: Polar coordinates can be useful for showing seasonal relationships in cyclical data with a fixed period (e.g. days in a week, months in a year).
Negatives: All else equal the outside group (disease) will look more prominent as the area will appear larger (even with square root scaling on the axis as done here). It can also be more difficult to see smaller groups relative to alternative chart types.
Comments: Could also be shown on line charts or (less preferably) a stacked bar/column chart.

Source: https://commons.wikimedia.org/wiki/File:Nightingale-mortality.jpg

Parallel Coordinates

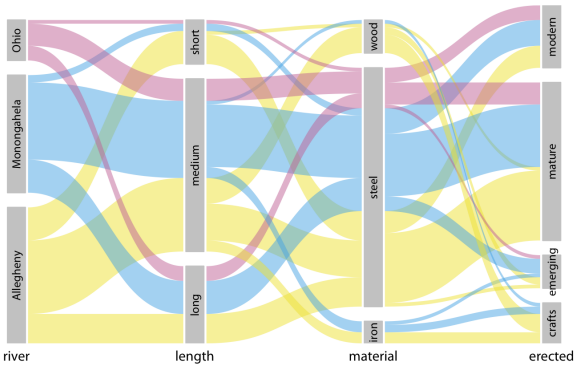


(c) Density-based parallel coordinates

Description: In a Parallel Coordinates Plot, each variable is given its own axis and all the axes are placed in parallel to each other. Each axis can have a different scale, as each variable works off a different unit of measurement, or all the axes can be normalised to keep all the scales uniform.
Data type: Multivariate, numerical data.
Positives: Parallel Coordinates Plots are ideal for comparing many variables together and seeing the relationships between them.
Negatives: The downside to Parallel Coordinates is that they can become over-cluttered when they're very data-dense.
Comments: The order the axes are arranged in can impact the way how the reader understands the data. One reason for this is that the relationships between adjacent variables are easier to perceive, then for non-adjacent variables. The example is atypical of a parallel coordinates plot, which would generally show lines between each parallel dimension.

Source: http://joules.de/files/heinrich_state_2013.pdf

Parallel Sets

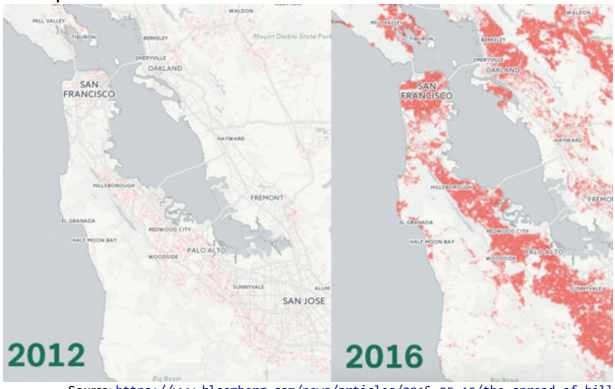


Description: Parallel Set charts are similar to Sankey Diagrams in the way they show flow and proportions. However, Parallel Sets don't use arrows and they divide the flow-path at each displayed line-set. Parallel set plots depict the proportional flow of information through a system. This can also be thought of as a sort of guide informing us of how the features in our data set are connected. The key advantage of Parallel set plots is the addition of the "proportional" component.
Data type: Categorical data showing different slices of a population.
Positives: Parallel sets are useful as they make ribbons of connection which are easy to spot.
Negatives: Can get messy with small categories. Data preprocessing can also be challenging.
Comments: May need to play with the ordering of the variables and the coloured variable to get a nice plot.

Source: https://clauswilke.com/dataviz/nested-proportions.html

Point Map

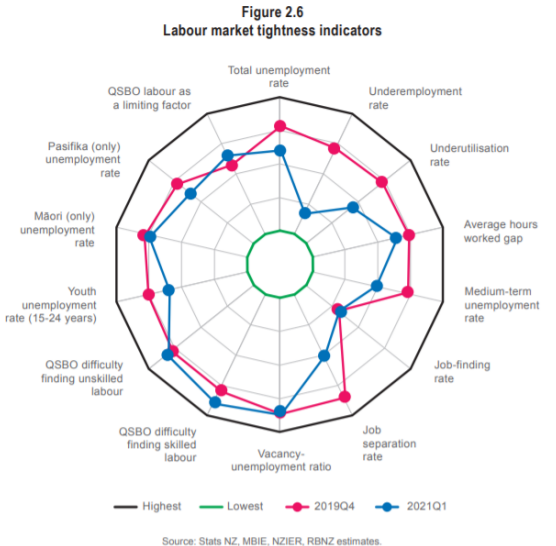
Spread of Million-Dollar Homes Across San Francisco



Description: Point Maps are a way of detecting spatial patterns or the distribution of data over a geographical region, by placing equally sized points over a geographical region.
Data type: Geospatial.
Positives: Intuitively shows the distribution of the points over a geographical region and can reveal patterns of where the points cluster on the map.
Negatives: Can be difficult to get exact values, particularly with a large number of points.
Comments: Essentially a scatterplot for geospatial data.

Source: https://www.bloomberg.com/news/articles/2016-05-19/the-spread-of-billionaire-s-bay-the-glut-of-million-dollar-homes-across-san-francisco

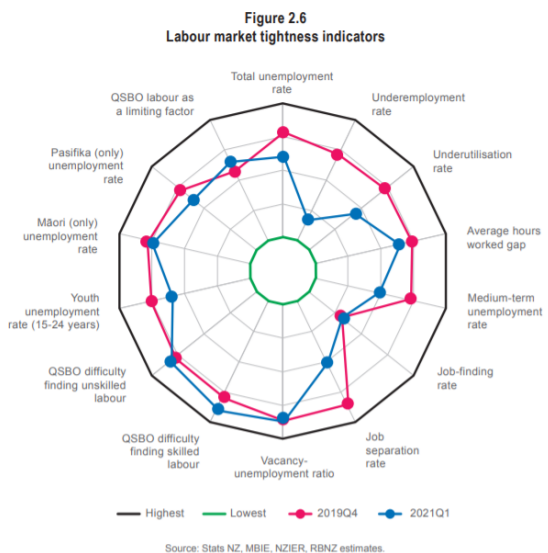
Radar Chart

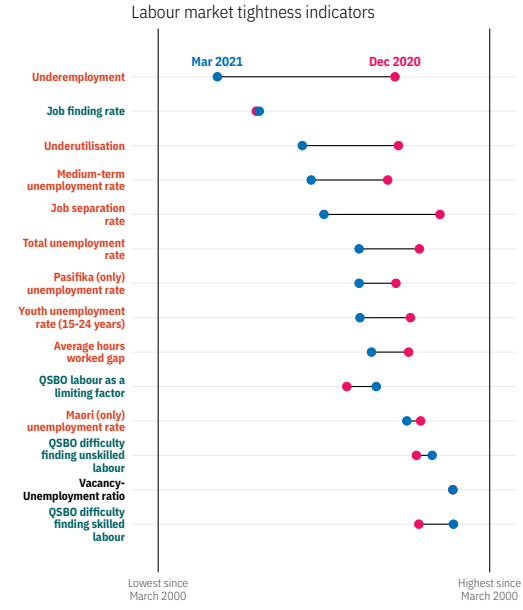


Description: A radar chart is a two-dimensional chart type designed to plot multiple series of values over multiple quantitative variables. Each variable has its own axis, all axes are joined in the center of the figure.
Data type: Multiple quantative variables, each with a numeric series.
Positives: Can show a lot of information in a small space.
Negatives: Category order can make the chart misleading. Circular layout is harder to read and compare. The area of a shape in a radar chart also increases quadratically rather than linearly, which could lead viewers to think that small changes are more significant than they actually are.
Comments: There are probably better alternatives…like the next slide.

Source: https://www.rbnz.govt.nz/monetary-policy/monetary-policy-statement/mps-may-2021

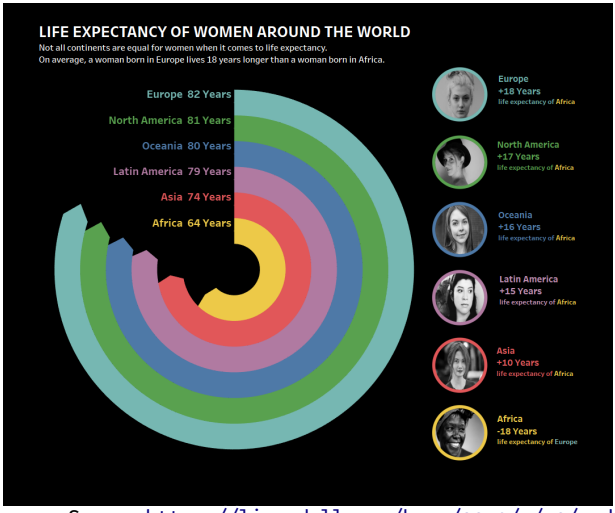
Radar Chart II





Source: https://www.rbnz.govt.nz/monetary-policy/monetary-policy-statement/mps-may-2021

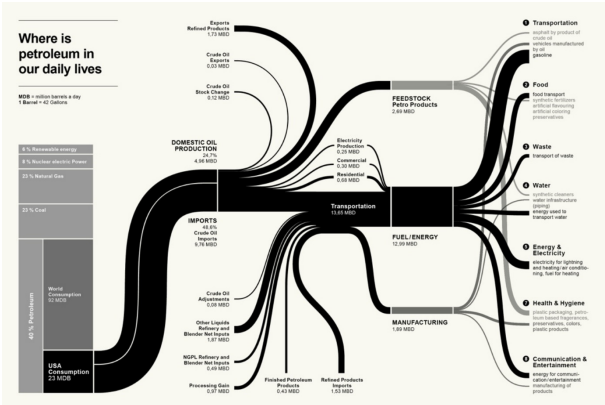
Radial Bar Chart



Description: A Radial/Circular Bar Chart is simply a Bar Chart plotted on a polar coordinate system, rather than on a Cartesian one.
Data type: Quantative categories, numeric values.
Positives: Visually distinctive.
Negatives: Because the bars are plotted on different radial points of the polar axis, they have different radii and cannot be compared by their lengths. A bar on the outside will be longer by construction than one on the inside, even with an equal value.
Comments: Generally better to use a bar chart.

Source: https://lisaadell.com/home/2019/7/10/radial-bar-chart-challenge

Sankey Diagram



Description: A sankey diagram is a visualization used to depict a flow from one set of values to another. The things being connected are called nodes and the connections are called links. Sankeys are best used when you want to show a many-to-many mapping between two domains or multiple paths through a set of stages.

Data type: Flows (Input/output) for a set of nodes.

Positives: Intuitive way of showing flows.

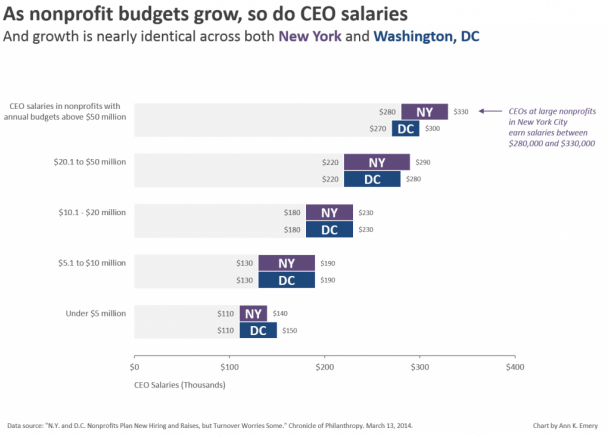
Negatives: Too many flows or flows which are too small can make it hard to interpret.

Comments: For the curious they're named after Captain Sankey, who created a diagram of steam engine efficiency that used arrows having widths proportional to heat loss.

Source:

https://www.ipoint-systems.com/blog/from-data-to-knowledge-the-power-of-elegant-sankey-diagrams/

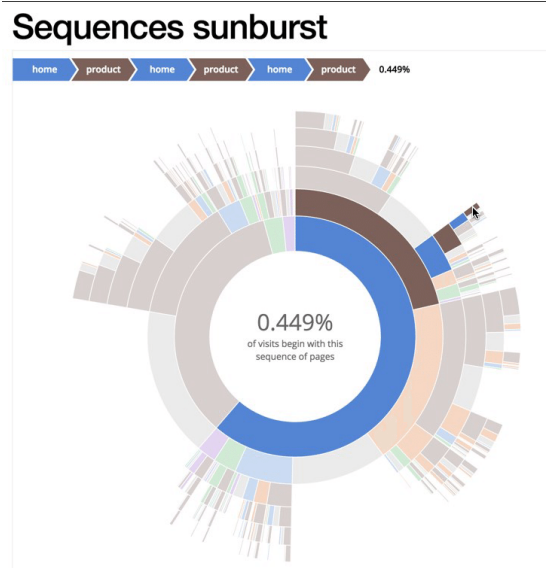
Span Chart



Description: A chart used to display dataset ranges between a minimum value and a maximum value.
Data type: Interval data, typically split into categories.
Positives: Span Charts are ideal for comparing ranges.
Negatives: Span Charts focus the reader on only the extreme values and give no information on the values in between the minimum and maximum values or on averages or data distribution.
Comments: Can also be used for timelines.

Source: https://depictdatastudio.com/span-charts/

Sunburst Diagram

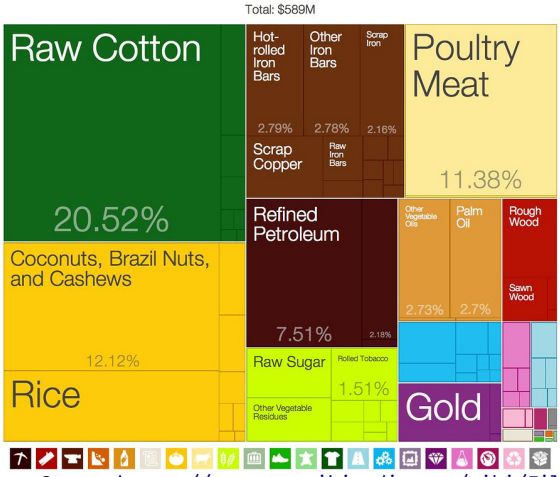


Description: This type of visualisation shows hierarchy through a series of rings, that are sliced for each category node. Each ring corresponds to a level in the hierarchy, with the central circle representing the root node and the hierarchy moving outwards from it.
Data type: Heirarchical.
Positives: Works well for data like website traversal (missing, repeats, etc.).
Negatives: Similar to other radial charts.
Comments: Earliest example goes back to William Playfair (1801).

Source: https://www.researchgate.net/figure/A-Sunburst-diagram-is-used-interactively-for-research-and-exploration-Brushing-across_fig3_303542683

Tree Map

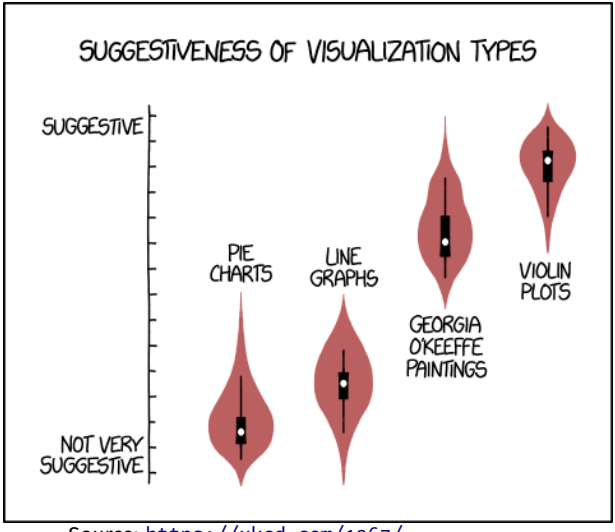
2009 Benin Exports



Description: Treemaps are a way of visualising the hierarchical structure while also displaying quantities for each category via area size. Each category is assigned a rectangle area with their subcategory rectangles nested inside of it.
Data type: Hierarchical with weights/quantities
Positives: Compact way of displaying hierarchical data, such as file sotrage on a computer.
Negatives: Can be difficult for non-specialists to interpret. Not very commonly used.
Comments: Simpler visualizations such as bar charts may be preferable.

Source: https://commons.wikimedia.org/wiki/File:Benin_English.png

Violin Plot



Source: https://xkcd.com/1967/

Description: A violin plot is essentially a mirrored kernel density estimator (fancy histogram) with a box plot overlaid.
Data type: Numerical, sometimes split into categories.
Positives: Presents a comprehensive distributional view in a compact space.
Negatives: Not very commonly used, and usually requires a lot of explanation.
Comments: Can be displayed either horizontally or vertically. Jittered data points can also be overlaid.

Mouseover text from XKCD: *Strictly speaking, 'violin' refers to the internal structure of the data. The external portion visible in the plot is called the 'viola'.*

Animation & interactivity

Colour & perception

Storytelling with Data – Practical Examples