



University
of Exeter

Data Visualisation

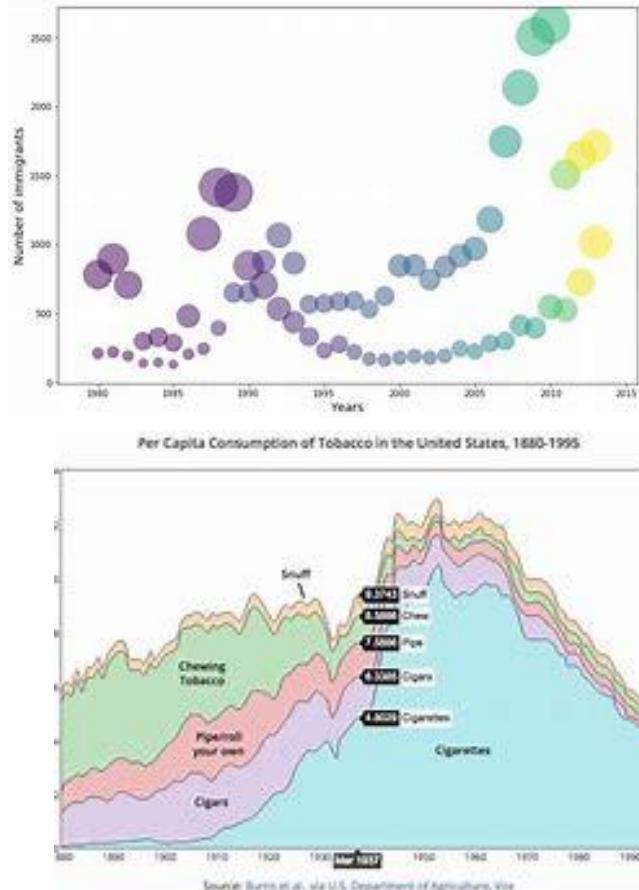
Week 03-BEM2031

Term2: 2023/24



Today:

- Assess the clarity of a visualisation (e.g. a graph)
- Understand various visualisation tools and what they are best suited for



Why visualize?

1. Exploratory Data Analysis
2. Communicating results



Histograms

Statistics

(summary statistics)

Descriptive
Statistics

Inferential
Statistics

Statistics

(summary statistics)

Descriptive Statistics

Central tendency

- Mean
- Median
- Mode

Dispersion / variability

- Range
- Interquartile range
- Variance
- Standard deviation

Skewness

- Symmetric
- Left
- Right

Inferential Statistics

Statistics

(summary statistics)

Descriptive Statistics

Central tendency

- Mean
- Median
- Mode

Dispersion / variability

- Range
- Interquartile range
- Variance
- Standard deviation

Skewness

- Symmetric
- Left
- Right

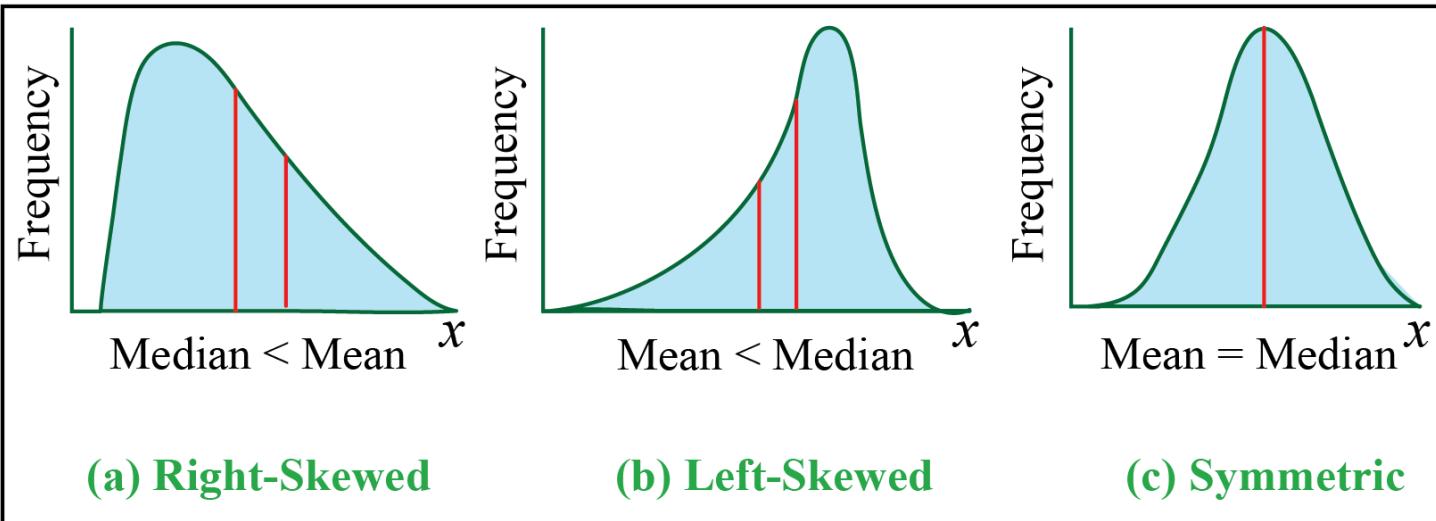
Inferential Statistics

Estimation parameters

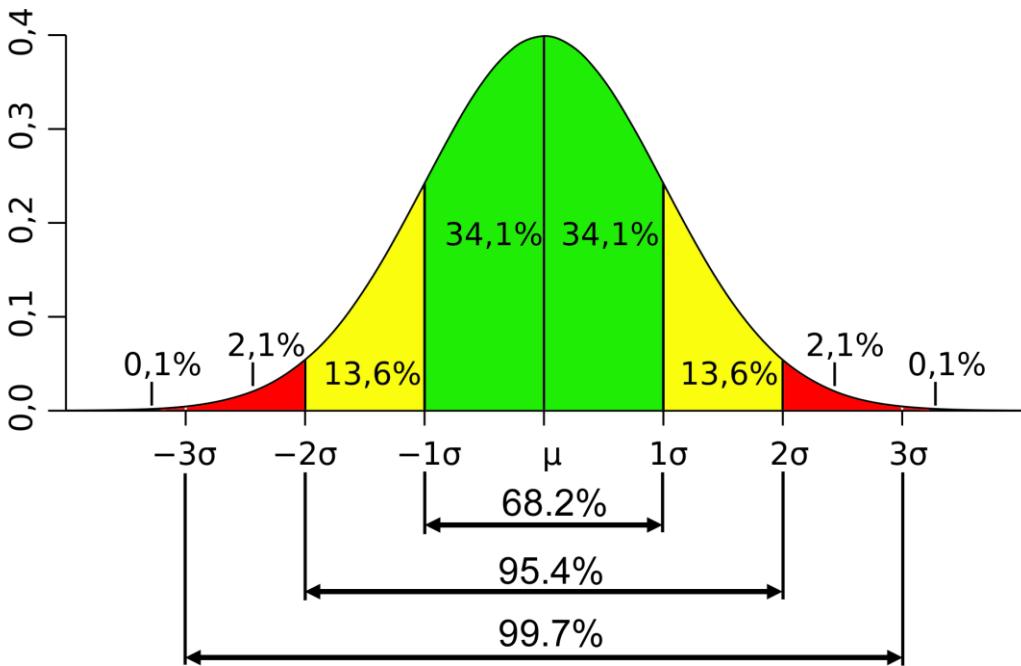
Hypotheses testing

Tests of difference and similarity

Statistics

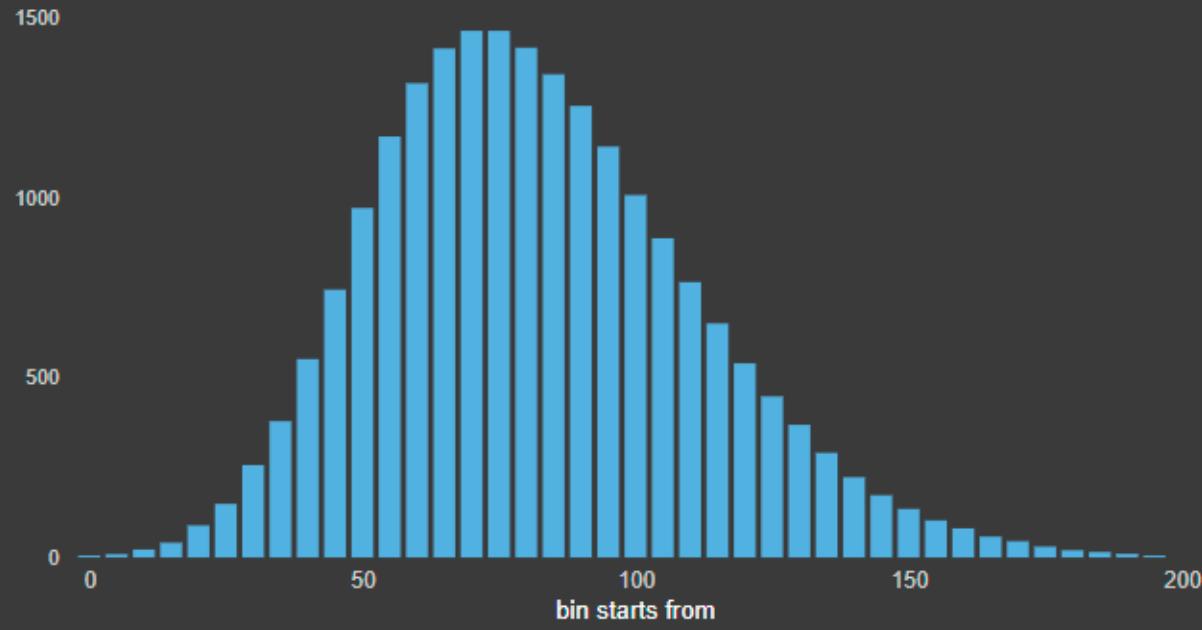


Statistics



Variance: the average squared differences from the mean

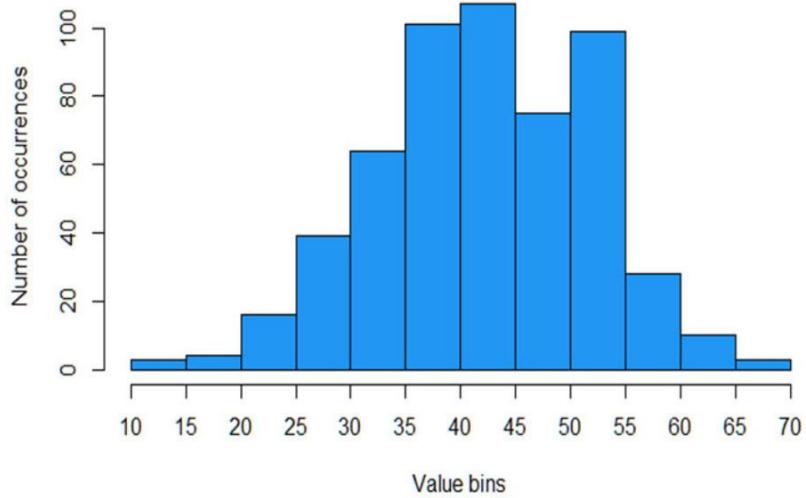
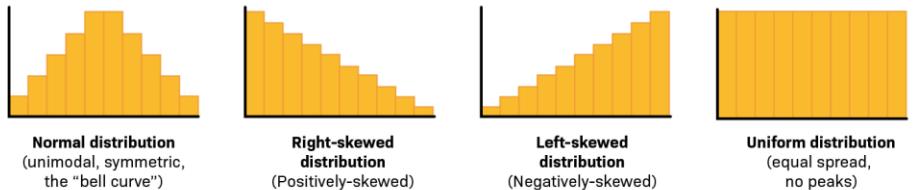
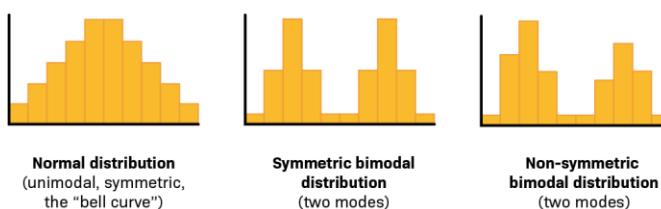
Standard deviation: the square root of the variance

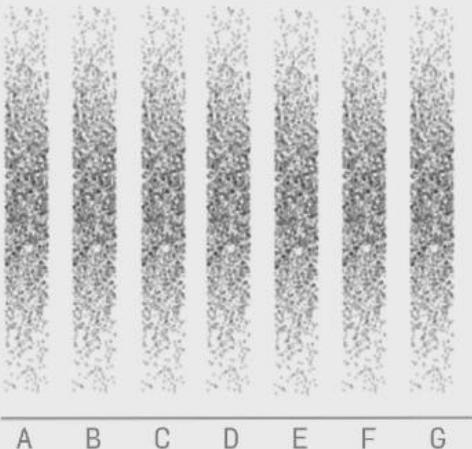
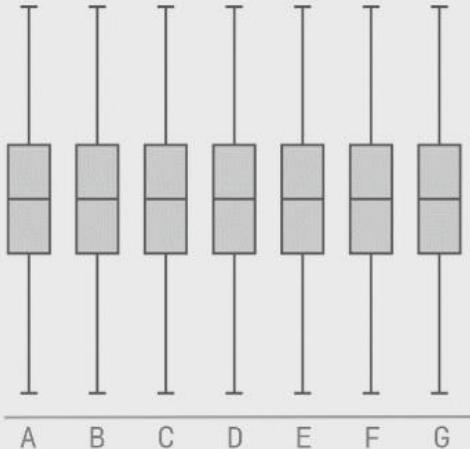
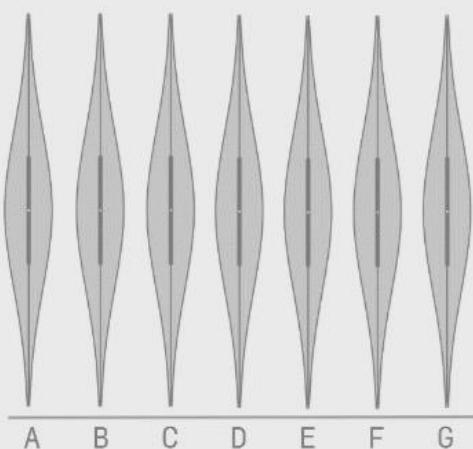


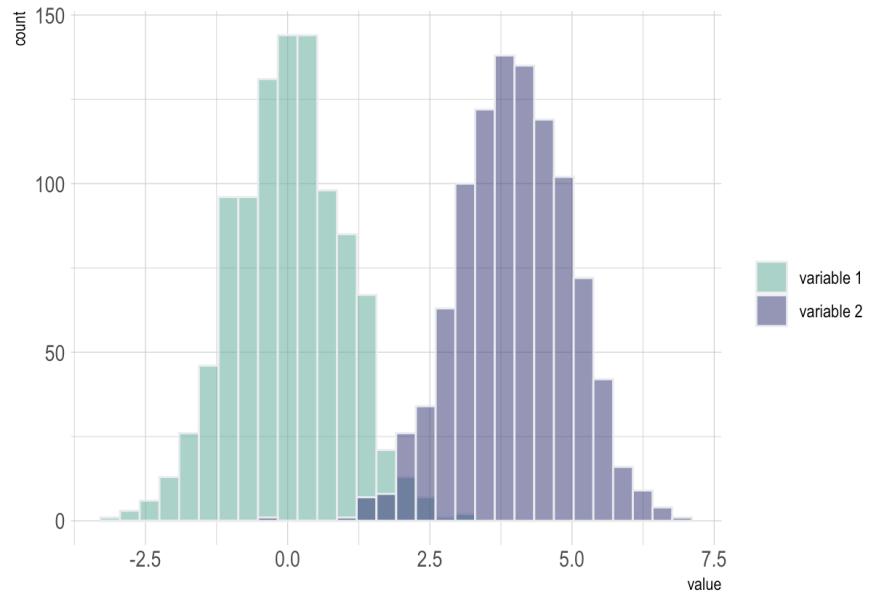
choose bin size

5

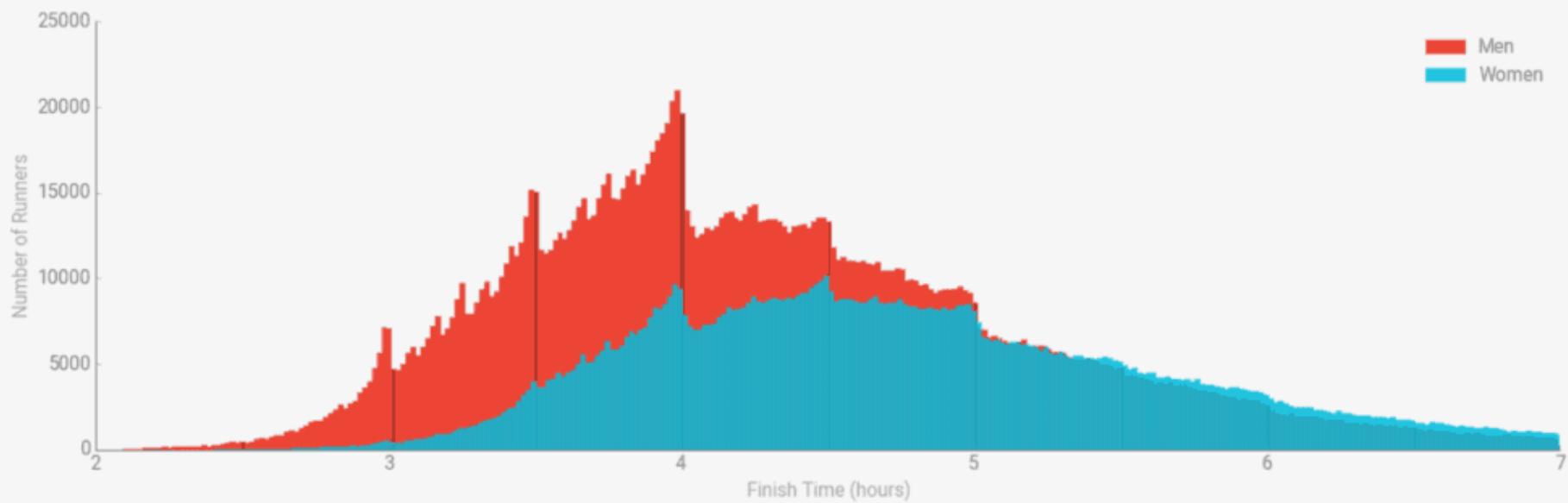


Symmetric (normal) vs skewed and uniform distributionsUnimodal vs bimodal distributions

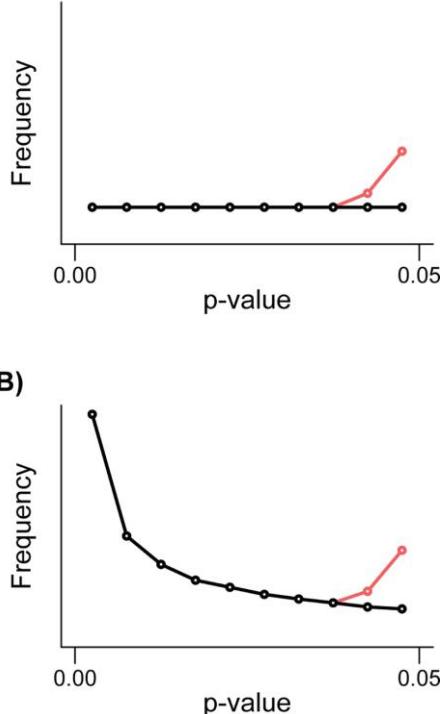
Raw Data**Box-plot of the Data****Violin-plot of the Data**



FINISH TIME DISTRIBUTIONS



A)

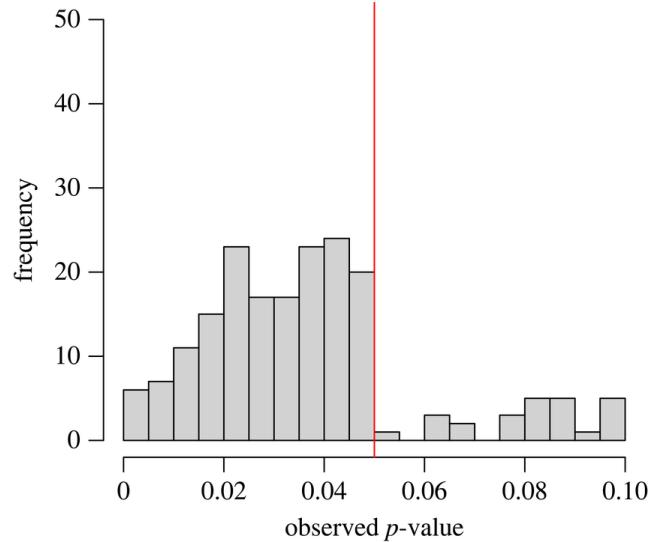
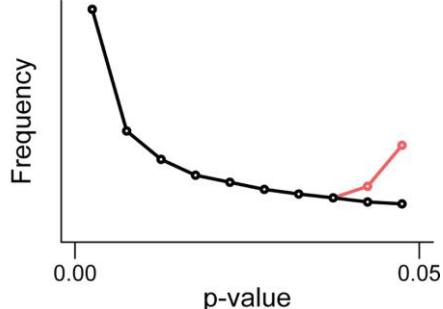


P-hacking is a bias in the scientific literature that occurs when researchers manipulate data or statistical analyses to obtain significant results.

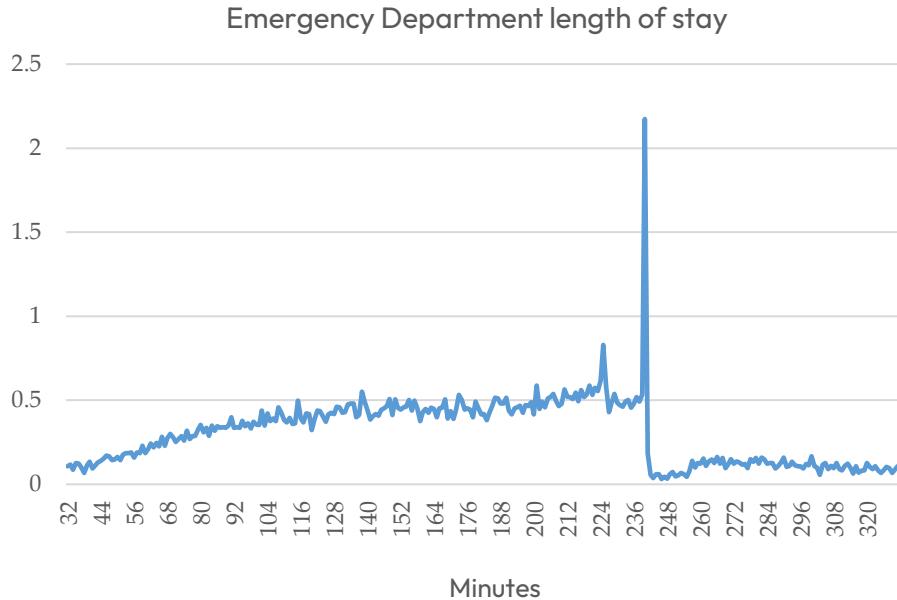
P-hacking can involve looking at many relationships, collecting or selecting data, or choosing different methods until non-significant results become significant.

P-hacking can lead to false or exaggerated findings and undermine the validity of scientific research.

B)

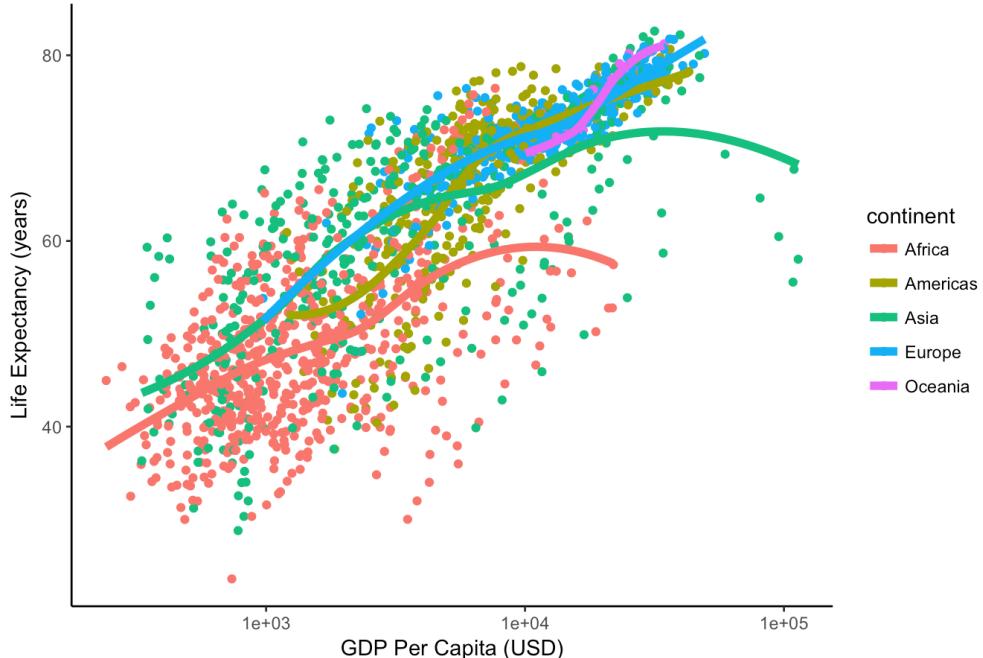


Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field. *Royal Society Open Science*, 9(12), 220946.



Can anyone
guess what is
happening in this
histogram?

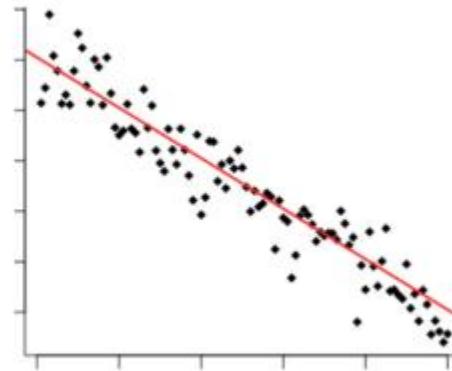
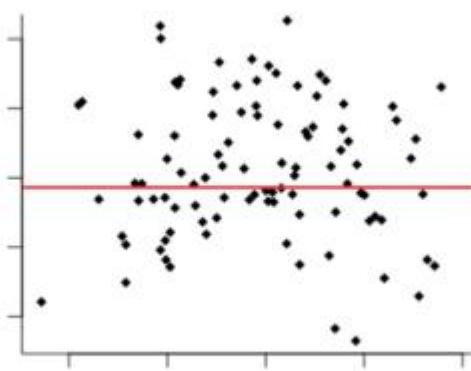
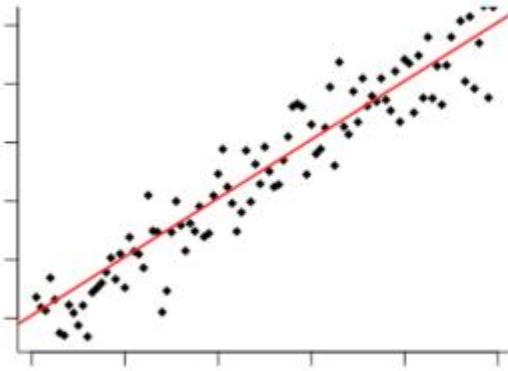
Life expectancy vs GDP by Continent



Scatterplots

Statistics

Correlation: the strength of a relationship between two variables. If two variables are correlated, as one changes in value, the other changes in the same direction.



Anscombe's quartet

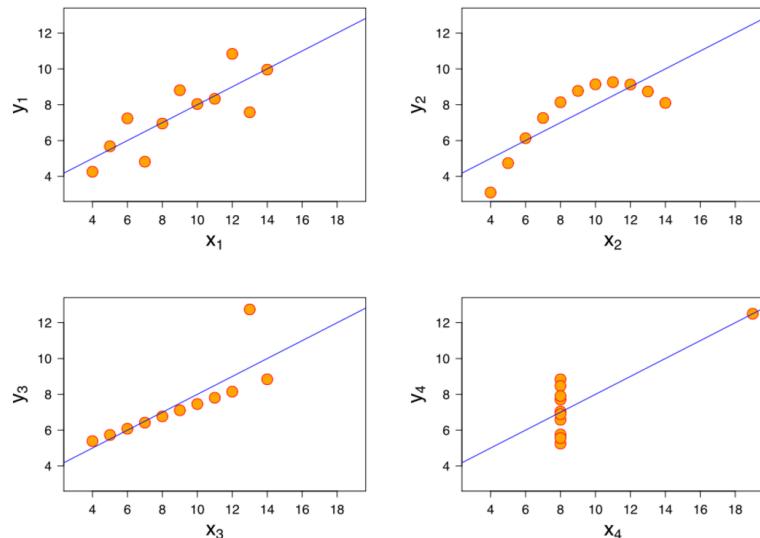
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

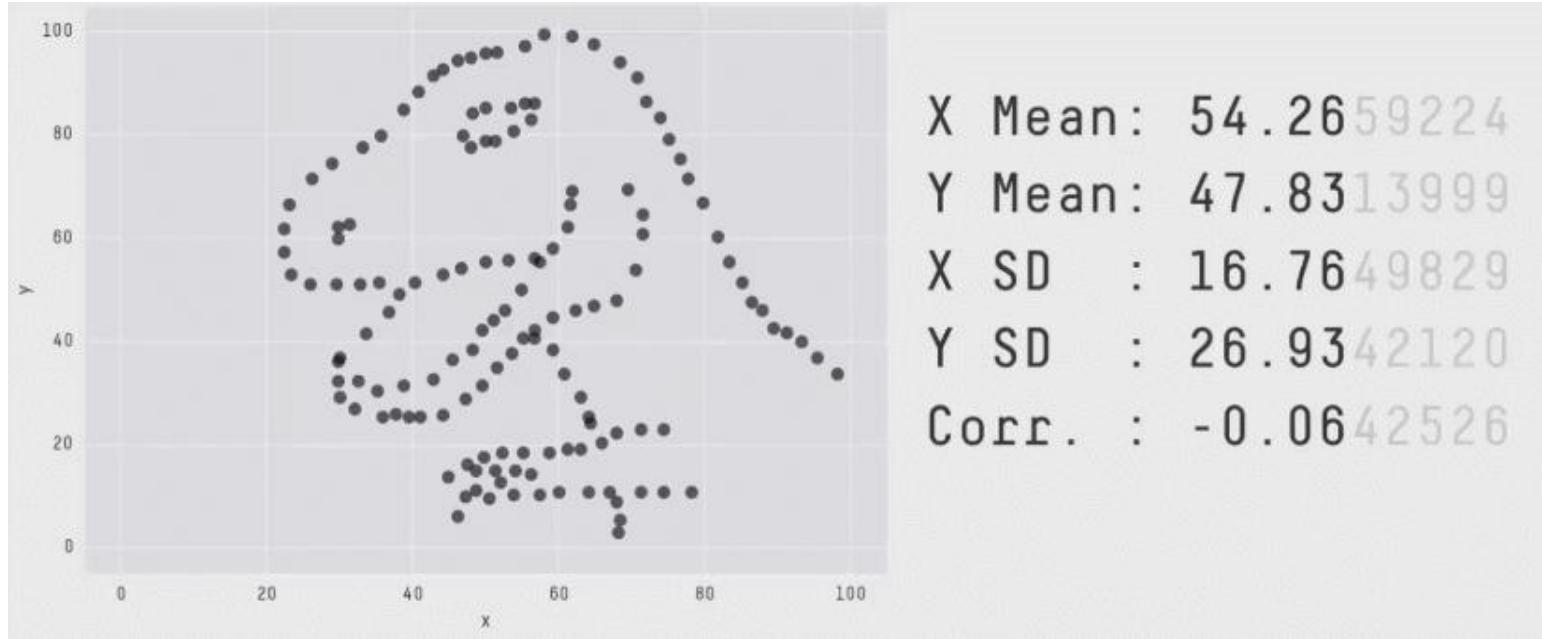
[Source](#)

Anscombe's quartet

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



"Same Stats, Different Graphs: Generating Datasets With Varied Appearance and Identical Statistics Through Simulated Annealing," by J. Matejka and G. Fitzmaurice, *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (<https://doi.org/10.1145/3025453.3025912>). Copyright 2017 Association for Computing Machinery.



"Same Stats, Different Graphs: Generating Datasets With Varied Appearance and Identical Statistics Through Simulated Annealing," by J. Matejka and G. Fitzmaurice, *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (<https://doi.org/10.1145/3025453.3025912>). Copyright 2017 Association for Computing Machinery.



Visual vocabulary

Visually Encoding Data

Cleveland, W. S., & McGill, R. 1984. *Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods*. *Journal of the American Statistical Association*, 79(387): 531–554.

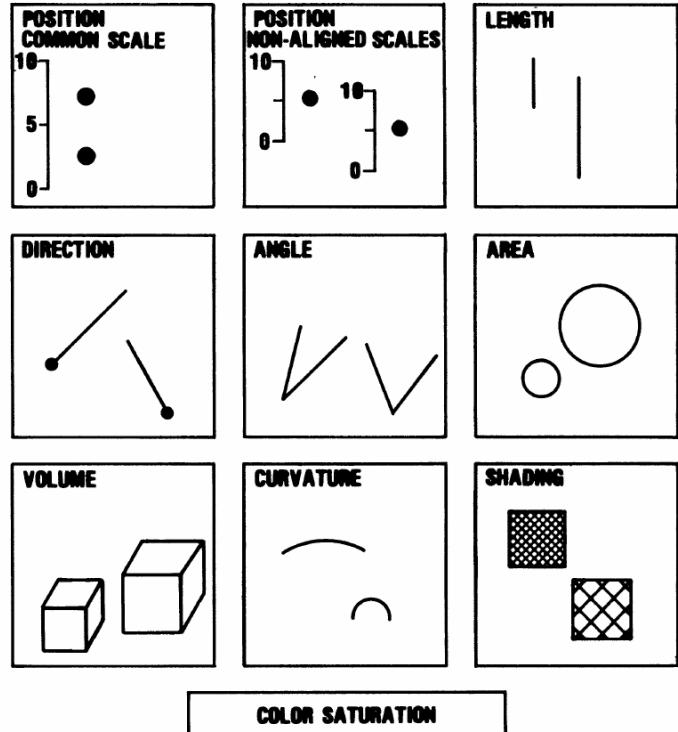


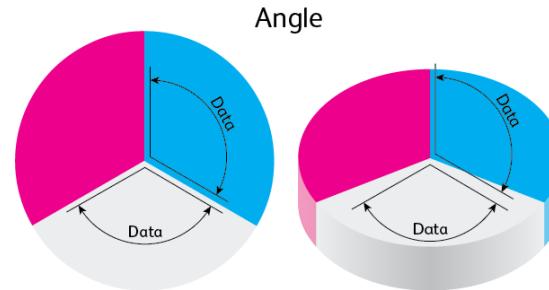
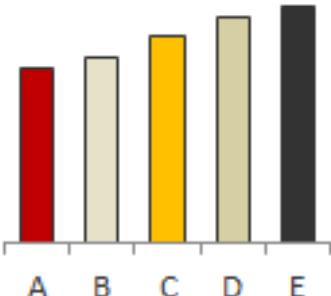
Figure 1. Elementary perceptual tasks.

Why do people hate Pie Charts?



University
of Exeter

Pie charts aren't *bad* visualizations. They just need to [be used appropriately](#).



The data in a pie chart is encoded in **the angle** of the slices.

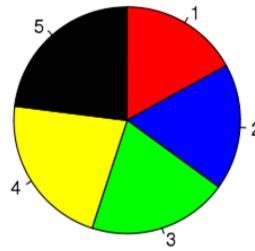
You may want to say it's encoded in the area, but if you create a pie chart by hand, what's the first thing you need to do?

Looking at Pie Charts....

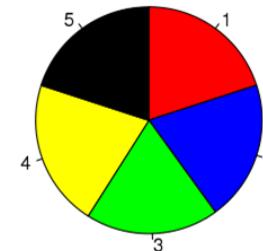


University
of Exeter

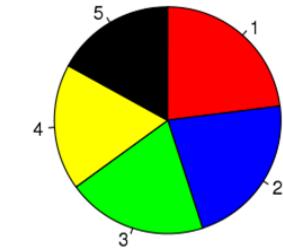
A



B



C

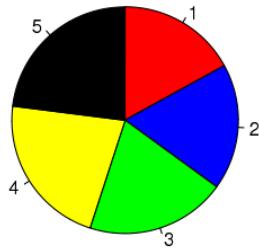


Looking at Pie Charts....

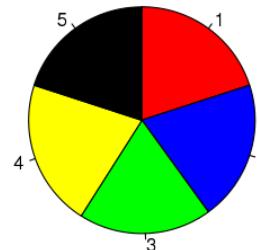


University
of Exeter

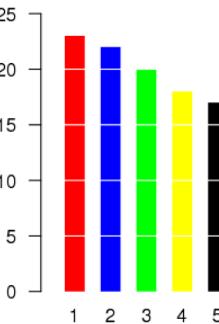
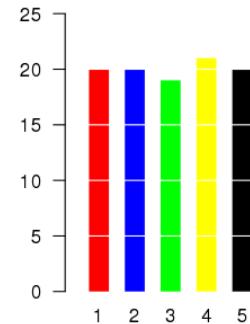
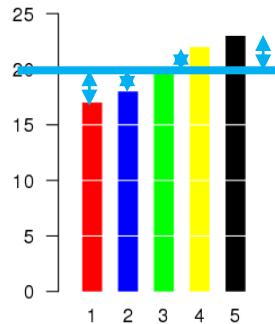
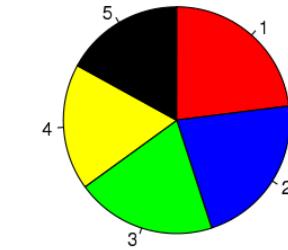
A



B



C

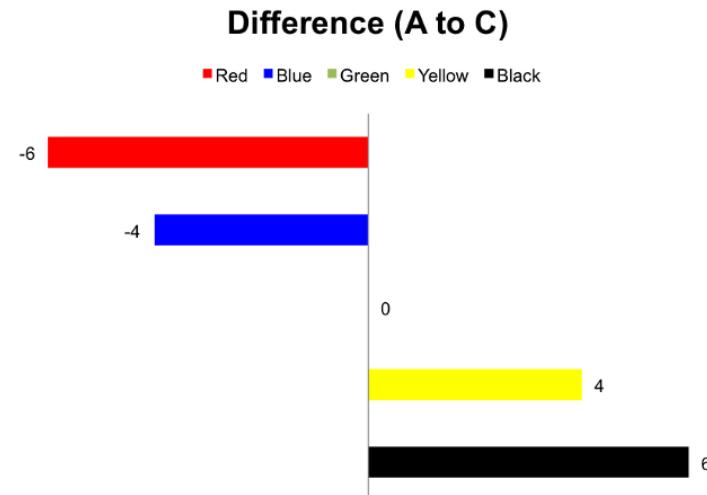


[Source](#)

Looking at Pie Charts....



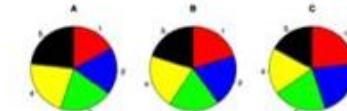
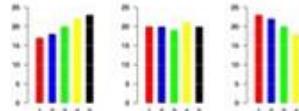
If you really need to compare the differences, then pie charts aren't what you need.



Looking at Pie Charts....



How important
is it to see
those small
differences?



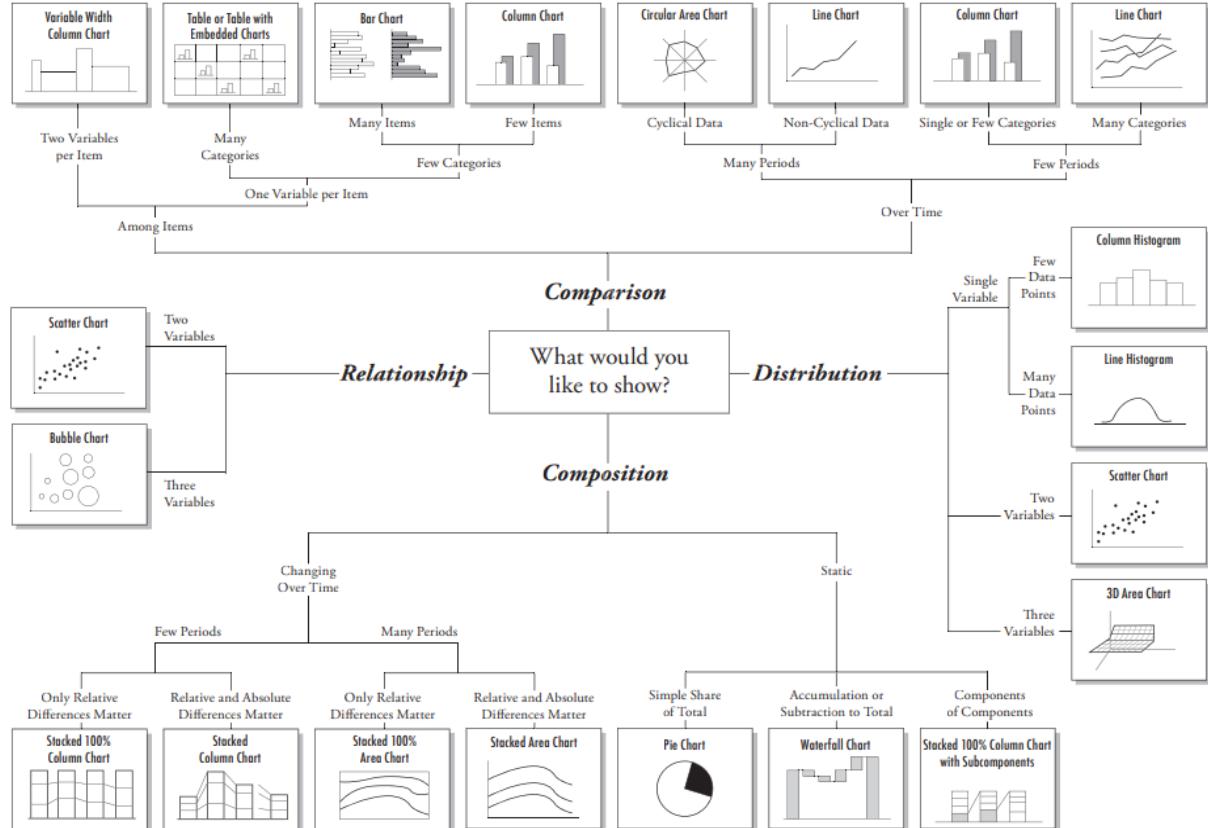
Easy to determine
relative differences:

*"Black changed more than
yellow"*

Easy to determine
absolute differences:

*"Neither yellow nor black
changed much"*

Chart Suggestions—A Thought-Starter





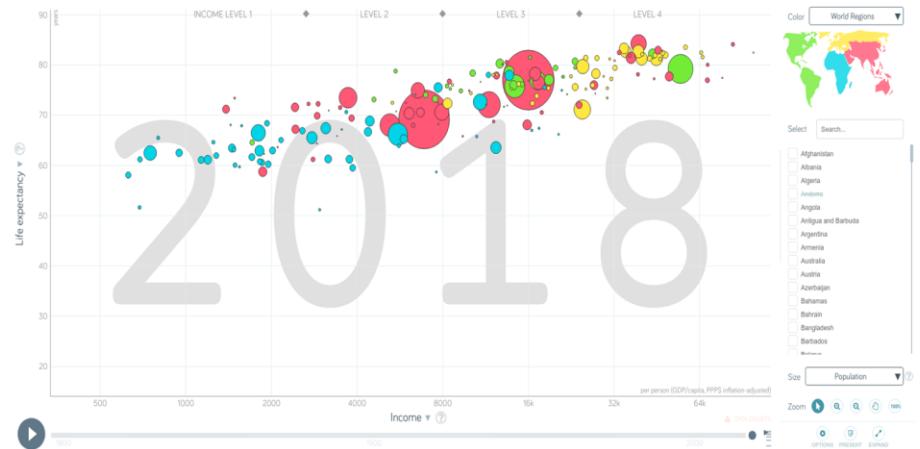
Deviation	Correlation	Ranking	Distribution	Change over Time	Magnitude	Part-to-whole	Spatial	Flow
Dotplots: variables (x) from a field point to point. Each point has a value. You tell them where, more modern tools will tell you what it is. Can be used to show something.	Show the relationship between two or more variables. If one variable goes up, does the other go up? If one variable goes down, does the other go down? Can be calculated (i.e. one counts the others).	Use where there's position in all dimensions. Correlation is positive or negative values. Don't tell me how far away they are, just tell me where he highlights his values. Values can be ranked.	Show values in a dataset and how often they occur. The easier the better if a right-skewed distribution. The level of diversity in the data.	One approach to changing trends. Trendline: a line that shows the movement in repeated events. Interpreting the current time period is important (the example bar chart is calculated one per cent).	Show size comparisons. That can be achieved by showing the area of the reader's eye. The reader's eye is static so the size of the area is constant.	Show how a single entity can be broken down into smaller entities. The reader's eye is static so the size of the area is constant.	Add from location maps only used for locations that have specific coordinates. That's why you might see a map of locations in the UK.	Show flat reader's eye. Only used for locations that have specific coordinates. That's why you might see a map of locations in the UK.
Example PT uses Taste, expenditure, income and the expenditure.	Example PT uses Wealth, deprivation, budget tables, communications results.	Example PT uses GDP, population, population (geographic distribution), recycling.	Example PT uses Dividend yields, market share, sales, seasonal changes in a market.	Example PT uses Communication, market capitalisation, volumes in general.	Example PT uses Face, budgets, company structures, regional, national, local, international.	Example PT uses Face, budgets, company structures, regional, national, local, international.	Example PT uses Physical objects, natural resources, locations, natural disaster, impact, boundaries, areas, variation in election results.	Example PT uses Physical objects, natural resources, locations, natural disaster, impact, boundaries, areas, variation in election results.
Giving bar 	Scatterplot 	Ordered bar 	Histogram 	Line 	Estates 	Stacked column/bar 	Bar 	Sankey
Giving stacked bar 	Ordered + line 	Ordered column 	Dot plot 	Dot plot 	Calms 	Basic bar/column/bar 	Bar chart 	Show changes in flows. Flows from one set of nodes to another, with varying widths for flow paths.
Name 	Connected scatterplot 	Ordered proportional symbol 	Dot skip plot 	Calms + line timeline 	Calms + line timeline 	Calms 	Paradox 	Chord
Surprise/filled line 	Bubble 	Column + line bubble 	Barcode plot 	Slope 	Calms 	Calms 	Flow map 	Network
XY heatmap 	XY heatmap 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Contour map 	Calms
Loftplot 	Slope 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Barplot 	Barplot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot 	Dot plot 	Violin plot 	Area chart 	Calms 	Calms 	Calms 	Calms
Violin plot 	Violin plot <img alt="Violin plot showing a distribution with a central box and a violin shape representing the spread of the							

The Hans Rosling Gapminder Tool

For a good example of compressing multiple dimensions of data into a single visualization, see [Hans Rosling's Gapminder tool](#).

Hans' talk is linked on the ELE page as well.

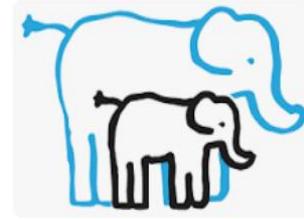
Gapminder has continued on without Hans' input since his passing and continued to help put world development and human experience data in context.



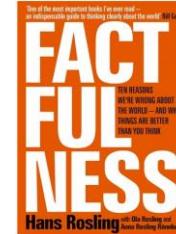


Principles of visualisations!

1. Clarity and Simplicity
2. Accuracy
3. Relevance
4. Consistency
5. Hierarchy and Emphasis
6. Effective Use of Colour
7. Labelling and Annotation
- 8 . Sufficiency
9. Interactivity
10. Storytelling
11. Chart Selection
12. Data-Ink Ratio
13. Audience Consideration
14. Accessibility
15. Testing and Feedback
16. Ethical Considerations
17. Credible Data Sources



Reference: Rosling 2019



Rosling, H. (2019). *Factfulness*, Flammarion. Available [electronically](#).

Principles of visualisations

1. Clarity and Simplicity: Keep the visualization simple and easy to understand. Avoid clutter, excessive decorations, and non-essential elements. The audience should be able to grasp the main message quickly.

2. Accuracy: Ensure that the data presented is accurate and that the visualization correctly represents the data. Misleading visualizations can harm the credibility of your information.

3. Relevance: Focus on the most important data and insights. Eliminate distractions and irrelevant details. Highlight what matters.

4. Consistency: Use consistent colours, scales, and terminology throughout the visualization. This helps the viewer make meaningful comparisons and understand the data more easily.

5. Hierarchy and Emphasis: Use visual hierarchy to guide the viewer's attention. Important elements should be more prominent, and less important elements should be de-emphasized.

6. Effective Use of Colour: Choose colours purposefully. Use colour to convey information, not just for decoration. Consider colourblind-friendly palettes. Too many colours can be confusing.

7. Labelling and Annotation: Clearly label data points, axes, and any relevant features. Annotations help provide context and explanations for the data.

8. Sufficiency: Provide enough data points to make the visualization informative but not overwhelming. Avoid overplotting, which can make the data hard to interpret.

9. Interactivity: For digital visualizations, interactivity can allow viewers to explore data in more detail. However, make sure it enhances understanding and doesn't create confusion.

10. Storytelling: Arrange data and visual elements in a logical sequence to tell a story. Help the viewer understand the narrative or insights you want to convey.

11. Chart Selection: Choose the right type of chart or graph for the data. Bar charts, line charts, pie charts, scatter plots, and others have different strengths for different types of data.

12. Data-Ink Ratio: Maximize the data-ink ratio, which is the proportion of ink (or pixels in digital formats) used to represent data compared to the total ink used in the visualization. Reduce unnecessary ink.

13. Audience Consideration: Understand your audience's background and familiarity with the subject matter. Adjust the complexity and terminology of the visualization to match the audience's level of expertise.

14. Accessibility: Ensure that your visualization is accessible to all, including individuals with disabilities. Use alt text for images, provide text descriptions, and follow accessibility guidelines.

15. Testing and Feedback: Test your visualization on potential users and gather feedback to make improvements. Different perspectives can help identify issues and areas for enhancement.

16. Ethical Considerations: Be mindful of the ethical implications of your data visualization, especially when dealing with sensitive or controversial topics. Avoid distorting or misrepresenting data.

17. Credible Data Sources: Clearly cite and reference the data sources used in your visualization to establish trustworthiness.





Chart Junk



Ink to Data Ratio

- The "data-ink ratio": Edward Tufte, a prominent expert in data visualisation.
- It refers to the proportion of ink (or pixels in digital formats) used in a visualisation that is directly related to representing the data, as opposed to ink used for labels, decorations, or non-essential elements.
- It encourages the minimisation of unnecessary ink to maximise efficient and clear visualisations, remove clutter, redundant, or distracting elements.
- It encourages designers to concisely convey the information the visualization is intended to communicate. This helps viewers quickly grasp the main message and insights from the data.

Relationship of Actual Rates of Registration to Predicted Rates
(104 cities 1960).

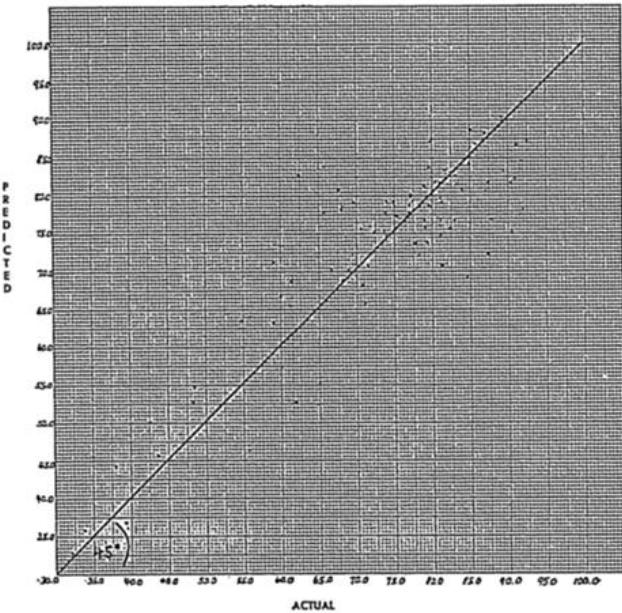
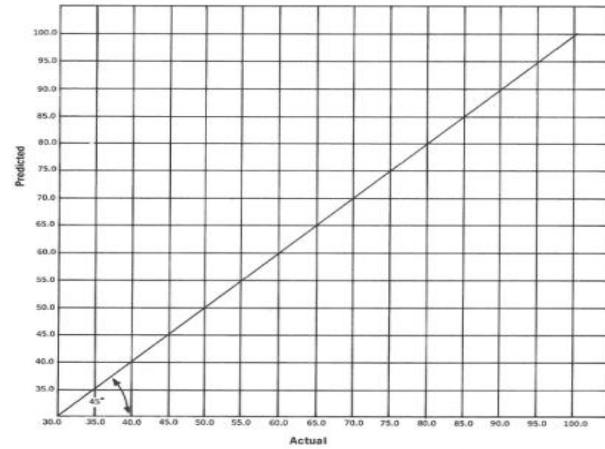


Figure 19.1 Relationship of Actual Rates of Registration to Predicted Rates
(104 cities, 1960)

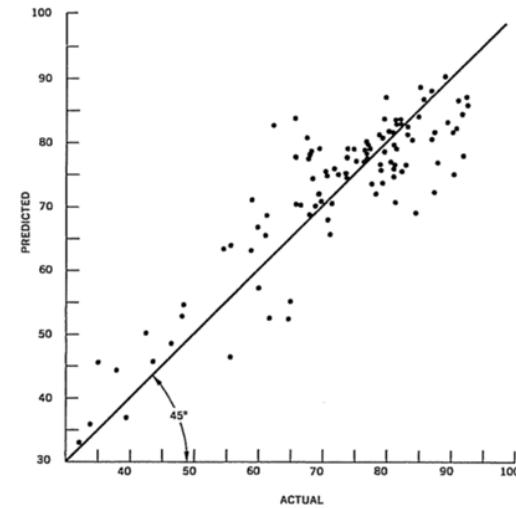


Principles of Data Ink

Above all else show data.

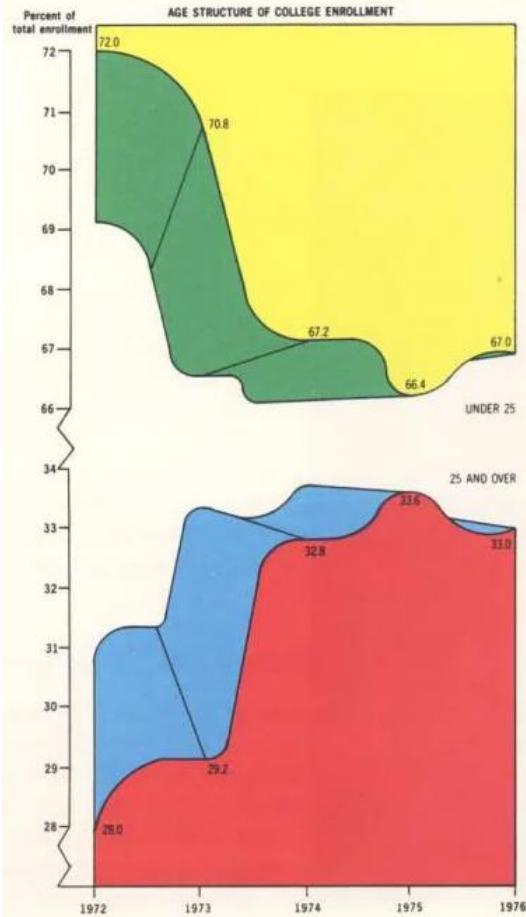


University
of Exeter



Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

Source: (Tufte, 2001)



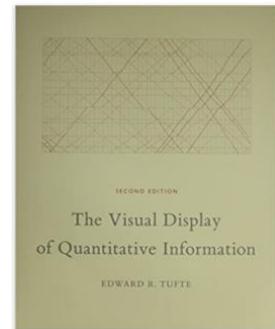
*The interior decoration of graphics generates a lot of ink that does not tell the viewer anything new... Regardless of its cause, it is all non-data-ink or redundant data-ink, and it is often **Chartjunk** (Tufte, 2001).*

Decorative effects, colouring, unnecessary graphical elements such as extra grid lines, redundant repetition of the data, excessive tick marks, etc can distract a viewer from understanding the underlying data shown in the graph.

Tufte (1983, p.118) says, "This may well be the worst graphic ever to find its way into print."

Excessive and unnecessary use of graphical effects – colour, 3D effects and disguised redundancy to represent just five numbers.

Tufte, E. R. (2001). *The visual display of quantitative information*, Graphics press Cheshire, CT.



Cool effects = Distorting data

What is the value for March? Can you tell?

There's some invisible tangent plane connecting to the "back" of the chart. These charts add an extra dimension, they add complexity, but there's no information in that dimension.

Number of issues

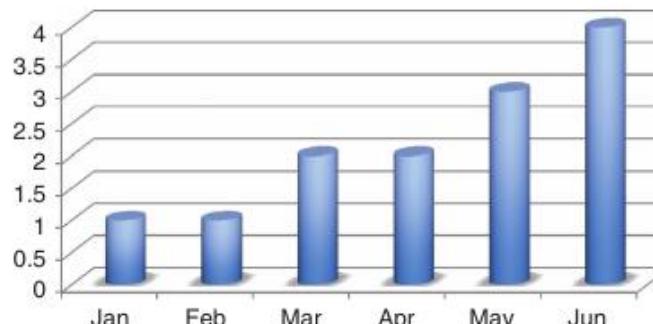
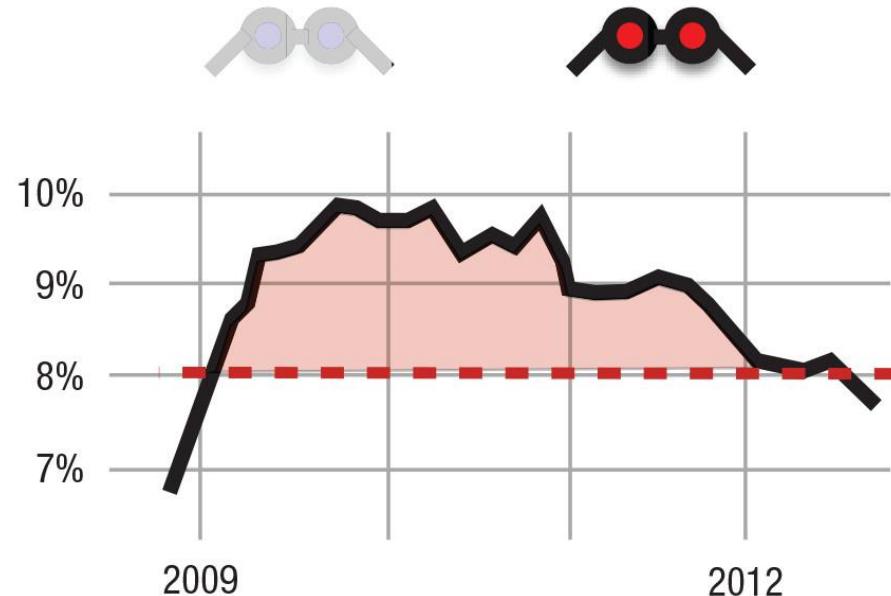
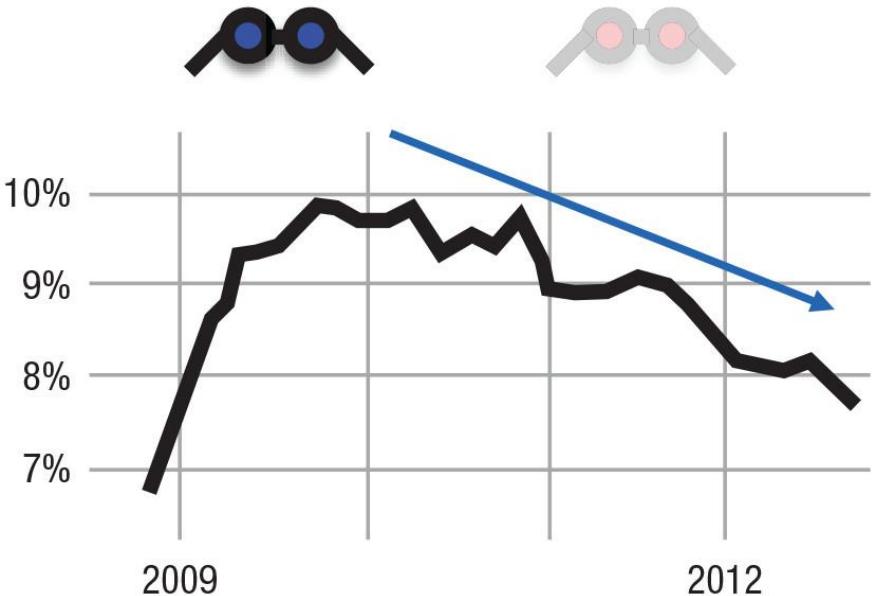


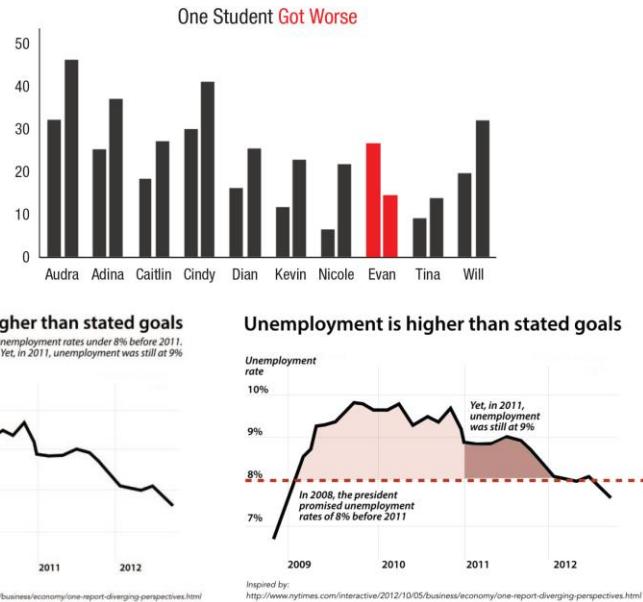
FIGURE 2.25 3D column chart



Remove
to improve
(the **data-ink** ratio)



An example of emphasizing different perspectives in a single data set (inspired by [Bostock et al., 2012](#)). One data set can be seen with dramatically different perspectives, depending on which patterns an observer does and does not extract.



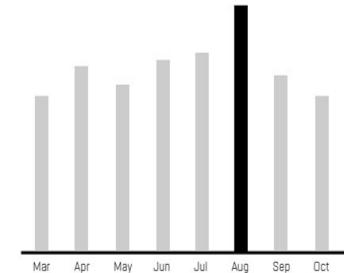
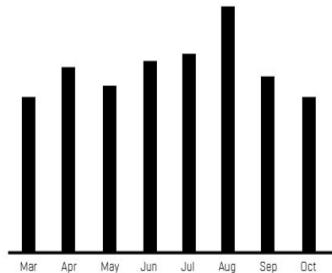
Colour highlighting and direct annotation to help viewers make the right comparison first and know what conclusion is supported by that pattern in the data.

The graphic at the top illustrates a colour-highlighting technique suggested in business-oriented practitioner guides (e.g., [Knafllic, 2015](#)).

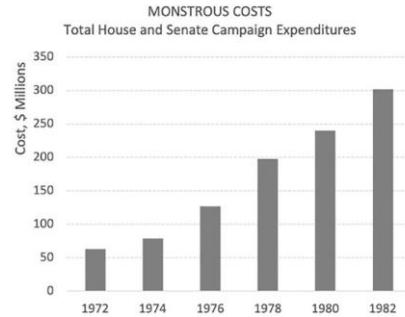
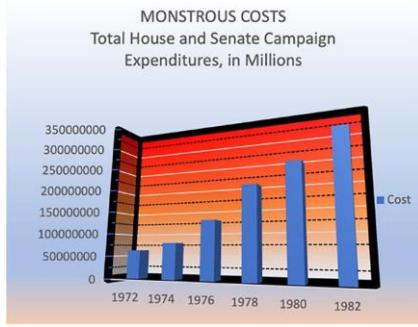
The graphs at the bottom (inspired by [Bostock et al., 2012](#)) are an adaptation of a graph by data journalists using grouping, highlighting and verbal annotation

The Squint Test

The squint test suggests you should blur your vision when looking at your visualization. After blurring your vision you should still be able to see some sort of pattern of interest.



[Image Source](#)



A “cluttered” visualization (top), a minimalist “decluttered” version (middle), and a version that incorporates pictorial embellishment (bottom).

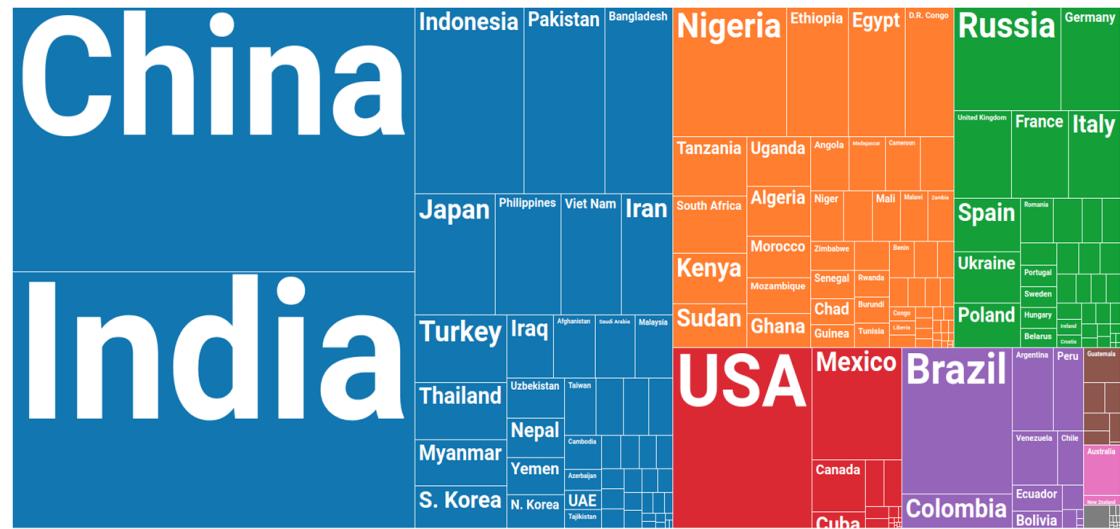
The graph at the bottom was created by Nigel Holmes for *TIME Magazine* and was reprinted in his [1984](#) book, *Designer's Guide to Creating Charts & Diagrams*.

Proportions. Tree maps.

Five alternatives to the pie chart.

Tree maps are a very useful way of visualization proportions. The data is *actually* encoded as area, and it's easier to break it up.

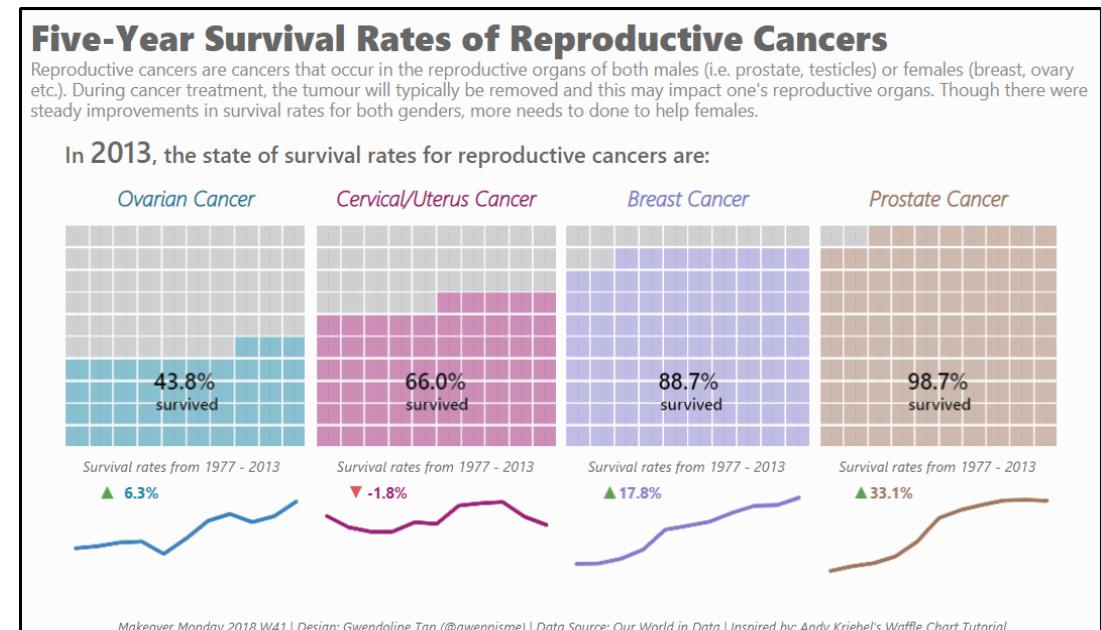
See this [interactive map](#) of population.



Proportions. Waffle plots.

Five alternatives to the pie chart.

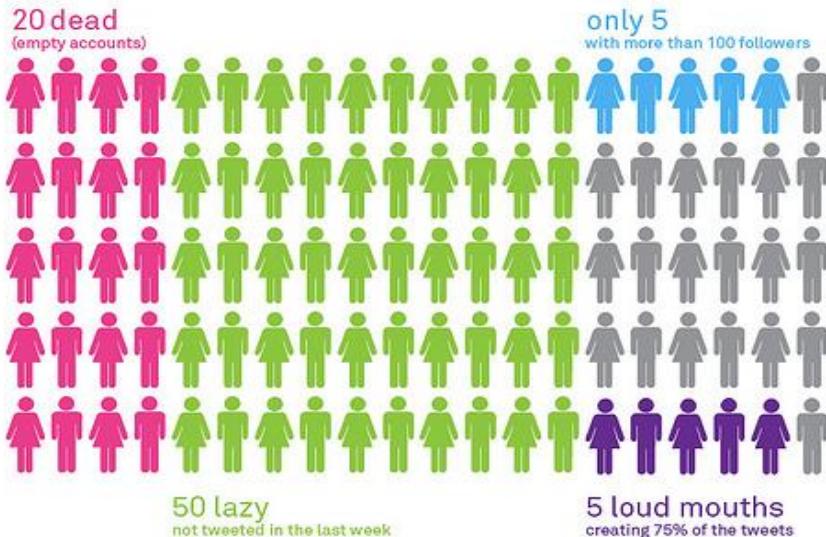
Waffle plots are great when you have a small amount of data. Or you just reduce a percentage to 100 boxes.



Proportions. Isotype/pictograms.

Reduce the population to 100 people. Then colour them proportionally.

Let's Not Get Too Excited...
If the Twitter community was 100 people...



David McCandless // www.visualizedthebook.com // v1.2

source: sysomos.com/insidetwitter/ [via rohitbhargava.typepad.com]

Proportions. Isotype/pictograms.

Reduce the population to 100 people. Then colour them proportionally.

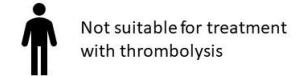
What problem are we addressing?

There is a gap between target thrombolysis (20%) and actual thrombolysis use (11-12%) in emergency stroke care

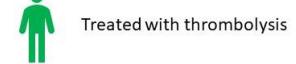
Clinical expert opinion on what *should be* happening



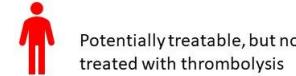
Unknown onset time or arrive too late to treat



Not suitable for treatment with thrombolysis



Treated with thrombolysis



Potentially treatable, but not treated with thrombolysis

What *is* happening?



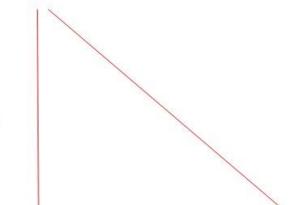
What did we test?

We used clinical pathway simulation and machine learning to analyze a series of 'what if?' questions:

1. What if arrival-to-treatment speed was 30 minutes?
2. What if all hospitals determined stroke onset time as frequently as the 'upper quartile' hospital (a hospital ranked 25 out of 100 hospitals)?
3. What if decisions were made according to a majority vote of 30 benchmark hospitals?

What did we find?

We found that making all these changes would increase thrombolysis use in England and Wales to 18-19%. Out of every 10 patients who were potentially treatable, but did not receive treatment, we found the cause to be:



Hospital processes were **too slow**

Stroke onset time was not determined when it potentially could have been

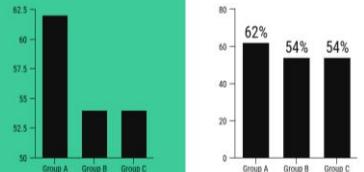
Doctors chose not to use thrombolysis when other higher-thrombolytic hospitals would have done



Using graphs to mislead

1 OMITTING THE BASELINE

In most cases, the baseline for a graph is 0. But writers can skew how data is perceived by making the baseline a different number. This is known as a "truncated graph".



MISLEADING

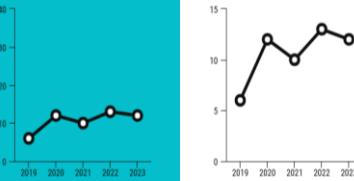
- Starting the vertical axis at 50 makes a small difference between groups seem massive.
- Group A looks much larger than Groups B and C.

ACCURATE ☺

- Starting the vertical axis at 0 offers a more accurate depiction of the data.
- The difference between the groups does not seem as dramatic.

2 MANIPULATING THE Y-AXIS

Expanding or compressing the scale on a graph can make changes in data seem more or less significant than they actually are.



MISLEADING

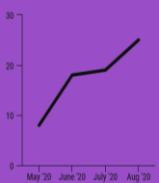
- The scale is disproportionate to the data, making the change over time seem small.

ACCURATE ☺

- The scale is proportionate to the data, showing a greater change over time.

3 CHERRY PICKING DATA

Writers may only include certain data points on their graphs to reinforce their narratives. This can create a false impression of the data.



MISLEADING

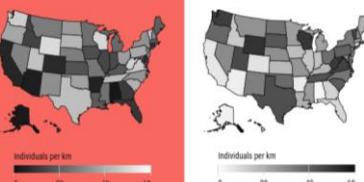
- Only a few months out of the year are graphed, depicting an upward trend.
- This graph shows the bigger picture.

ACCURATE ☺

- A much wider date range is graphed, revealing an overall downward trend.
- This graph shows the bigger picture.

5 GOING AGAINST CONVENTIONS

Over time, we have developed standards for how data is visualized. Flipping those conventions can make a graph confusing or misleading to readers.



MISLEADING

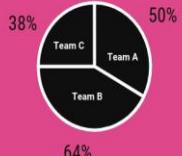
- Normally, darker shades are associated with density on a map but here, dark has been used to depict lower population density.

ACCURATE ☺

- This map follows the convention of using lighter shades for lighter density and darker shades for higher density.
- Readers will intuitively know how to interpret the data.

4 USING THE WRONG GRAPH

The type of graph you use should depend on the type of data you want to visualize. Using the wrong type of graph can skew the data. Writers will sometimes use the wrong type of graph on purpose.



MISLEADING

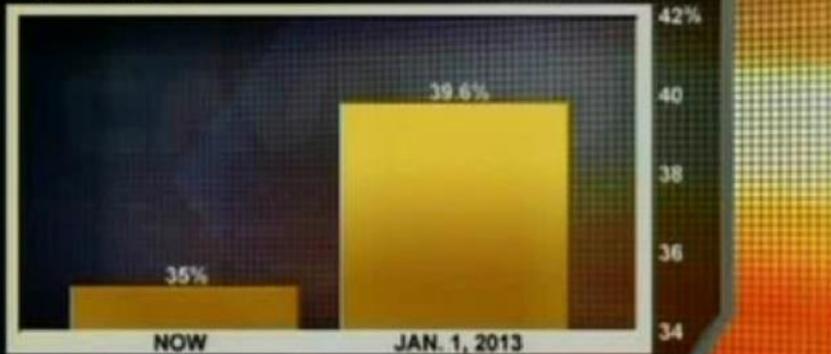
- Pie charts are used to compare parts of a whole, not the difference between groups.
- A different type of graph should be used to compare the three teams.

ACCURATE ☺

- Bar graphs are better for showing the differences between groups.
- This chart is a better visualization of the data.

IF BUSH TAX CUTS EXPIRE

TOP TAX RATE



8:01p ET

FOX
BUSINESS

TOP STORIES

TECHNOLOGY

CONSUMER

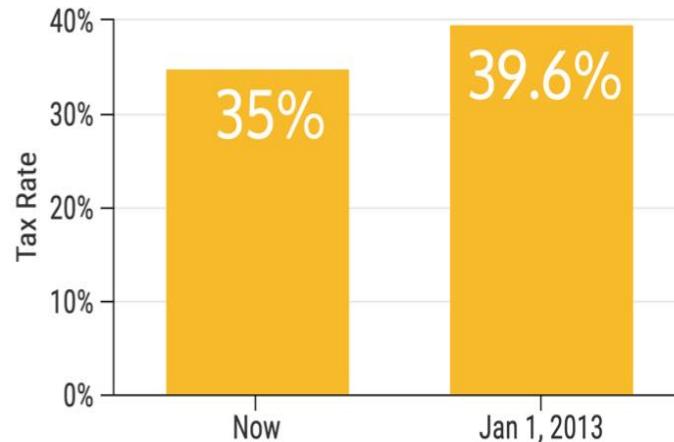
WITH THE JUSTICE DEPARTMENT AND ACQUIRES FULL T

DOW 13008.68 □ 64.33

S&P 1379.32 □ 5.98

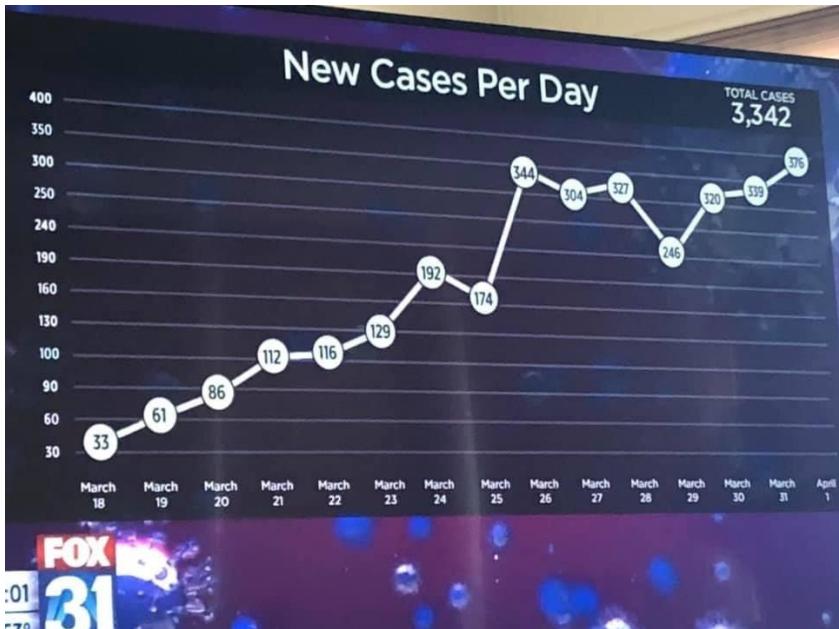
NASDAQ 2939.52 □ 6.32

If Bush Tax Cuts Expire



Deceptive Designs

Deceptive Design



The gridlines are equally spaced on the page, but sometimes the same space represents 30 people, sometimes 10, and sometimes 50.

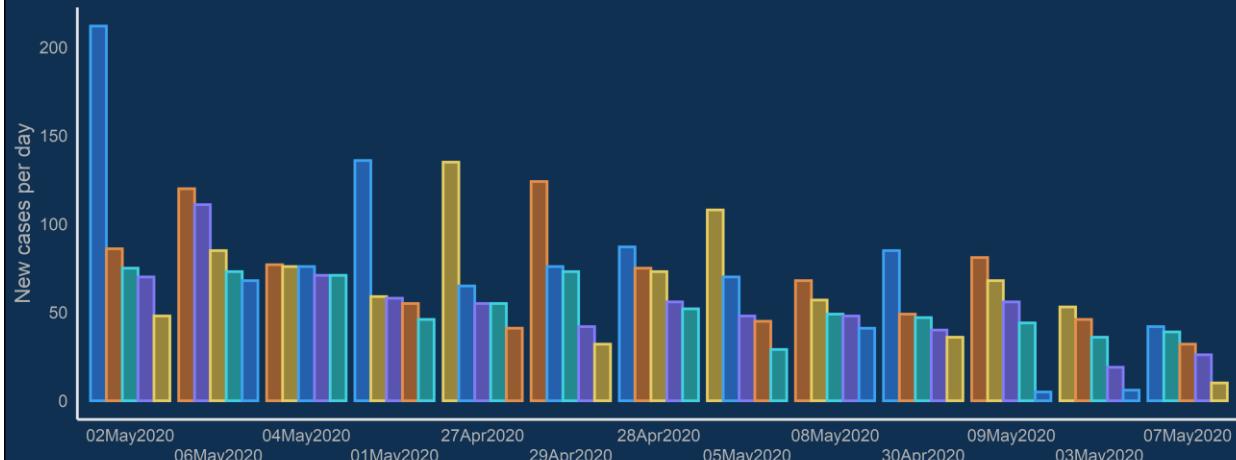
It isn't even strictly increasing - which might be expected if someone was trying to crudely copy by hand the effect of a logarithmic scale - but seems to be completely arbitrary.

Deceptive Design

Top 5 Counties in Georgia with the Greatest Number of Confirmed COVID-19 Cases

Note that this chart is to illustrate poor visual design choices and does not include the most current data. It uses different data from the original from the Georgia Department of Public Health.

█ Hall
 █ Gwinnett
 █ Fulton
 █ Cobb
 █ DeKalb



Source: analysis by <http://freerangestats.info> with county-level COVID-19 case data from New York Times

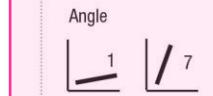
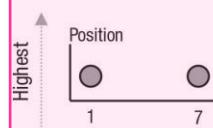
As a visualisation at least three things are wrong:

1. Dodged bar charts are rarely effective for making comparisons over time - it's difficult for the eye to follow;
2. Within each day's clump of bars, the counties are in a different order (highest to lowest, within the clump), reducing the meaning in the pattern in each clump;
3. The daily clumps of bars are not in chronological order.

[Ordering bars within their clumps in a bar chart \(freerangestats.info\)](#)

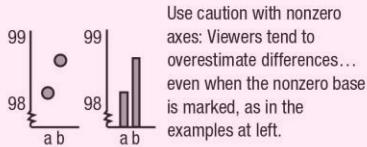
Absolute Precision Ranking for Seeing a Single Ratio

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

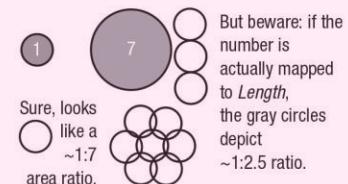
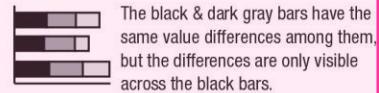


Common Illusions That Distort Data

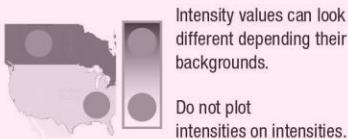
Caveats for the visual encoding in each row



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

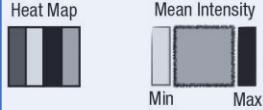
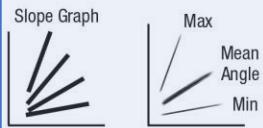
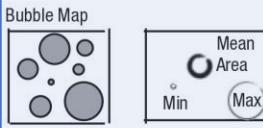
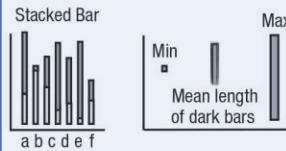
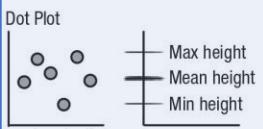


The difference is larger for the lighter segments compared with the darker ones, right? That is an illusion—the differences are identical.



Vision Is Powerful for Global Statistics

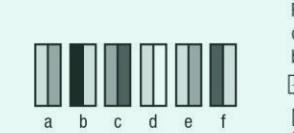
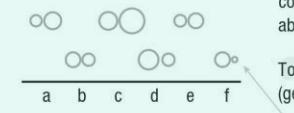
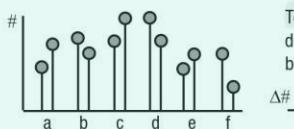
For each visualization, statistics are available quickly



Vision Is Sluggish for Comparisons

Isolating pairs with “larger second values” is tough...

So guide viewers to the right comparisons



The two columns on the left show a quick reference guide to channels that can depict data visually and common illusions for each channel.

The column in the centre presents a summary of how visual statistics are powerful.

The two columns on the right illustrate how comparisons are severely limited and present a set of design techniques that focus viewers on the “right” ones.

Next Week: Clusters and Similarity

-  • Read Data Science for Business, chapter 6 (book available [electronically](#)).
-  • Watch: StatQuest: [K-Means clustering](#)
-  • Watch: StatQuest: [Principal Component Analysis \(PCA\) Step-by-Step](#)
-  • Watch: StatQuest: [PCA Main Ideas](#)
-  • Play: [Visualizing K-Means Clustering](#)
-  • Play: [Visualizing DBSCAN](#)
-  • Play: [Principal Component Analysis](#)



University
of Exeter



Any questions?

?