



University
of Exeter

Data Visualisation

Week 03-BEM2031

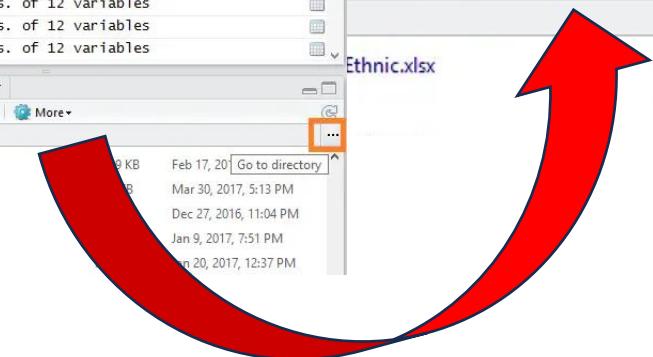
Term2: 2024/25



Today:

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

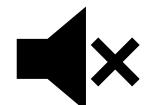
- Most queries I have had are related to not setting the working directory. R needs to know where to look for a file, such as a csv, if asked to access one.



The screenshot shows the RStudio interface with several windows open:

- Code Editor:** Shows R code for generating data frames and performing data manipulation.
- Environment:** Shows the global environment with objects like A, agg, agg_mean, agg_sum, B, C, clust1, clust2, and clust3.
- Viewer:** Shows the results of running the R code, including data summary tables for clust1, clust2, and clust3.
- File Explorer:** Shows a file tree with files like Ethnic.xlsx, Only womens Ethnic.xlsx, pgadmin.log, brand_fashion_dim.sas, RENT RECEIPI_2016-2017.docx, and seller Analysis.sas.

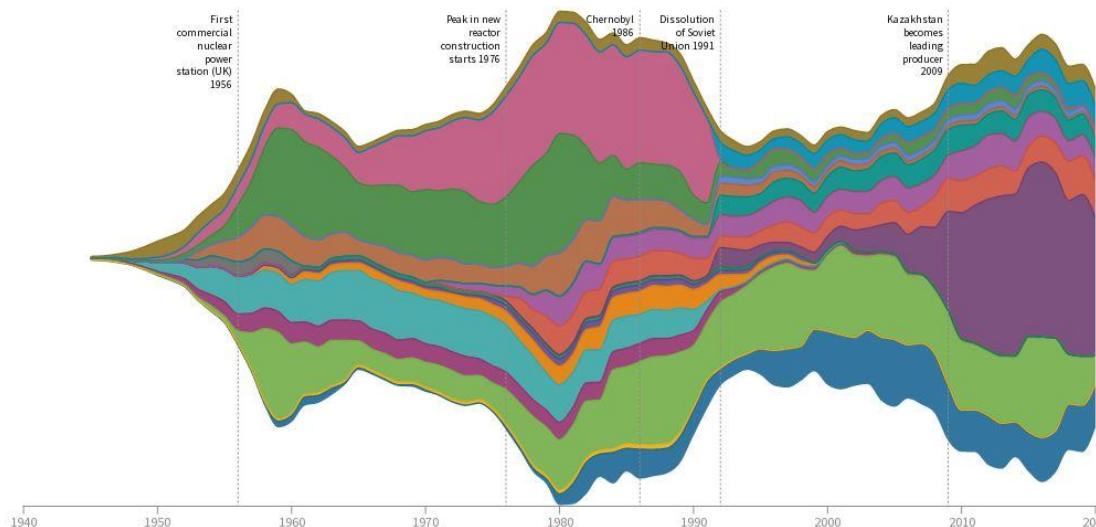
- Make sure you can knit to pdf and/or html from your R markdown files. If you are having problems knitting to pdf you may need to use **install.packages('tinytex')**
- I have had several requests to please QUIETEN DOWN in workshop labs – it is hard for everyone to concentrate with the amount of noise in the room!!



World uranium production, 1945-2020

Units are tonnes uranium

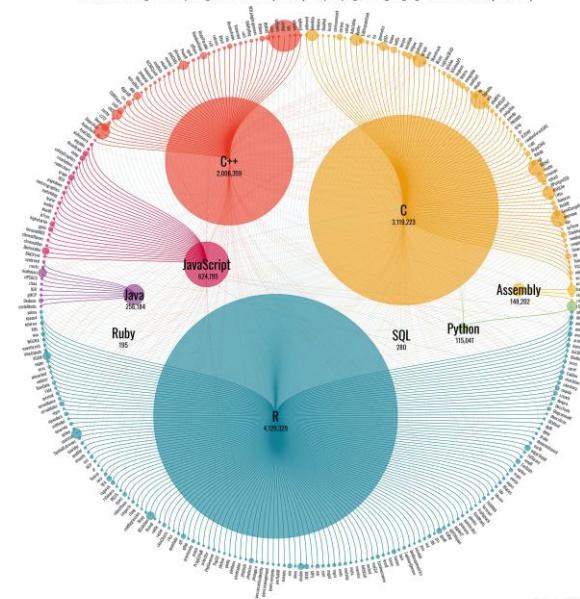
All



Source: OECD Nuclear Energy Agency
Visualizing Energy | Institute for Global Sustainability | Boston University

LOC of Popular Programming Languages in 300 CRAN Packages

considered are largest CRAN packages written in one (or more) of top 16 programming languages from Tobe Index (Nov. 2019)



#Wistia H3C236 gender

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

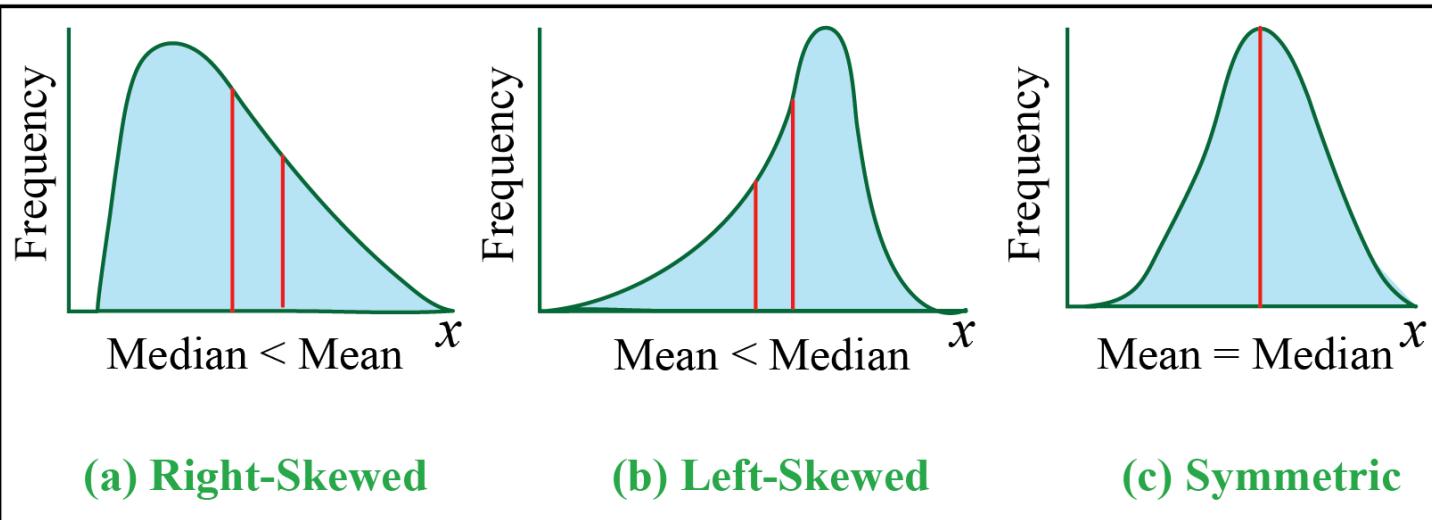
Inferential Statistics

Estimation parameters

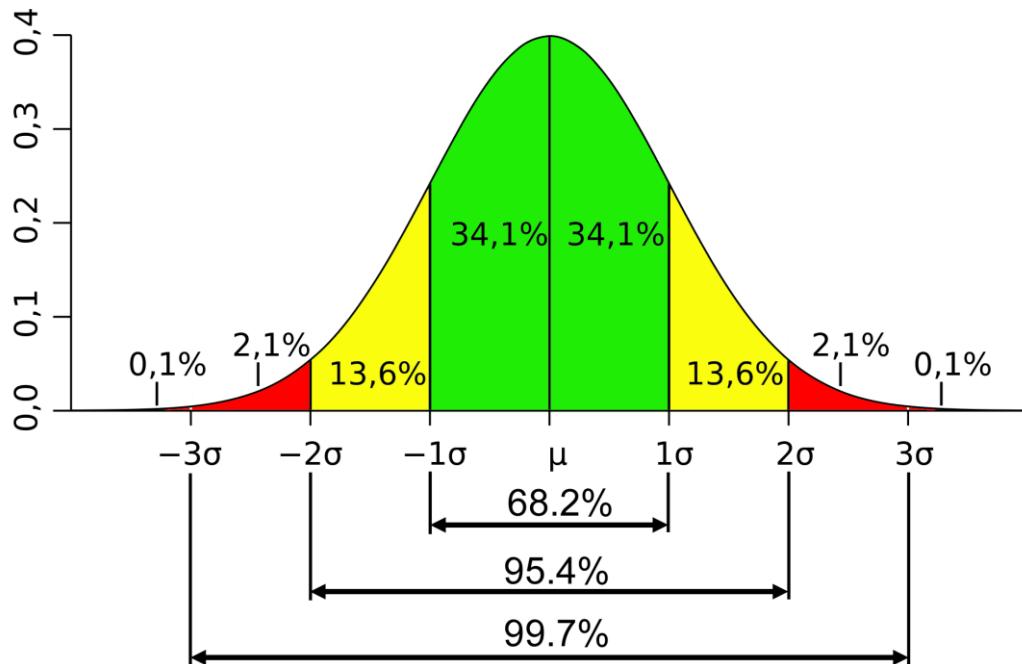
Hypotheses testing

Tests of difference and similarity

Statistics

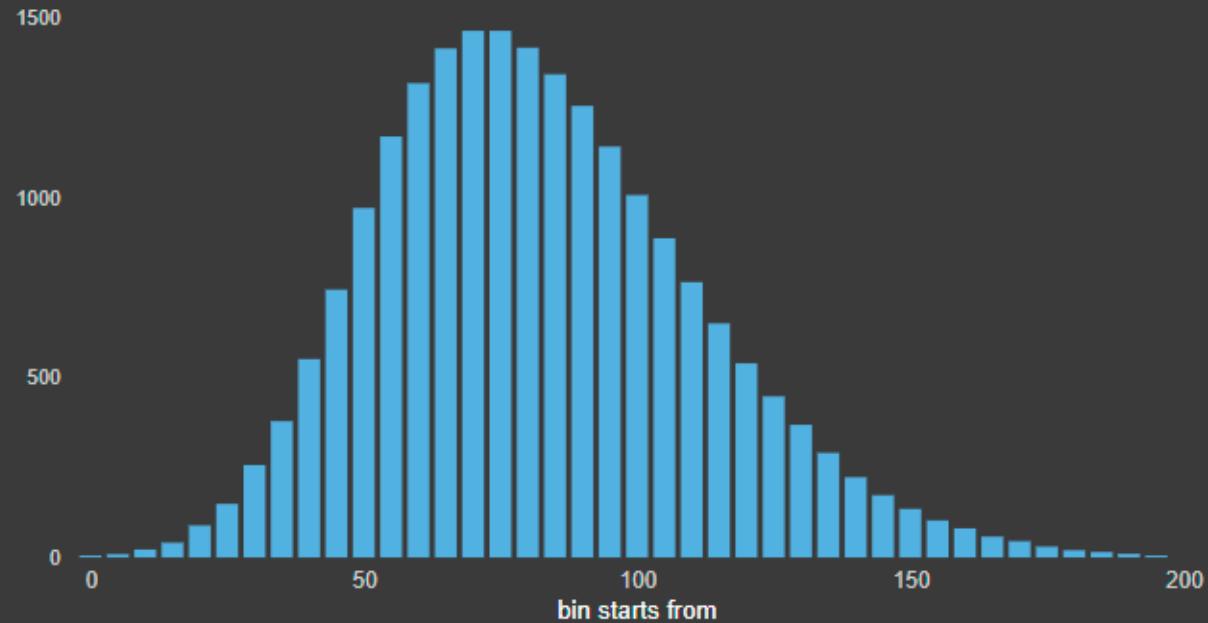


Statistics



Variance: the average squared differences from the mean

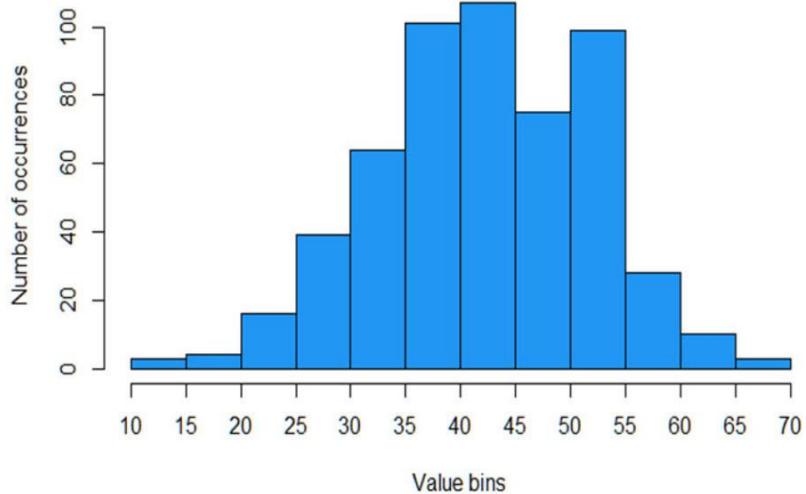
Standard deviation: the square root of the variance



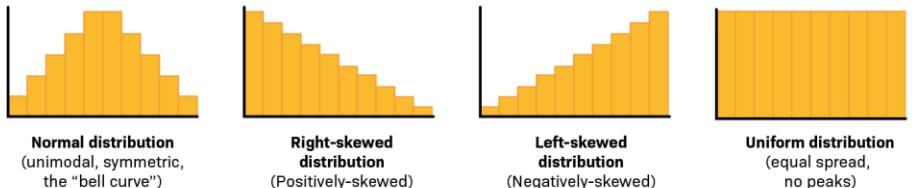
choose bin size

5

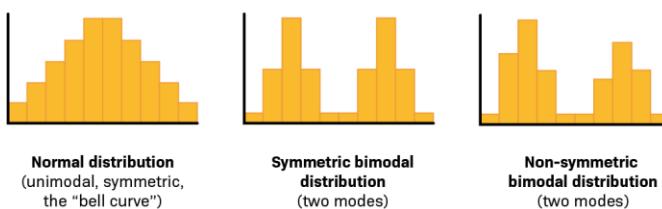


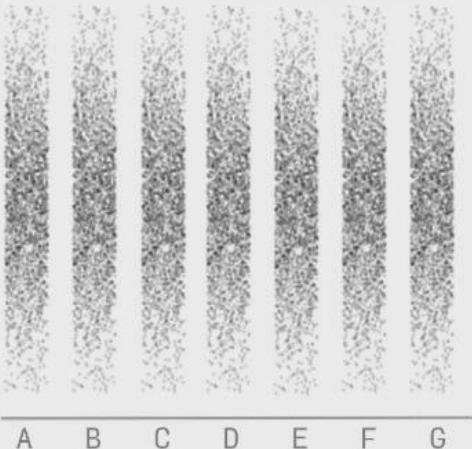
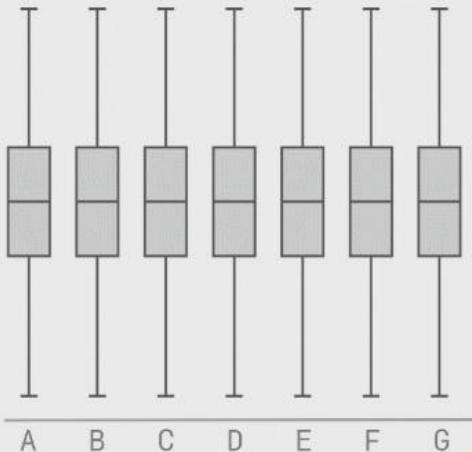
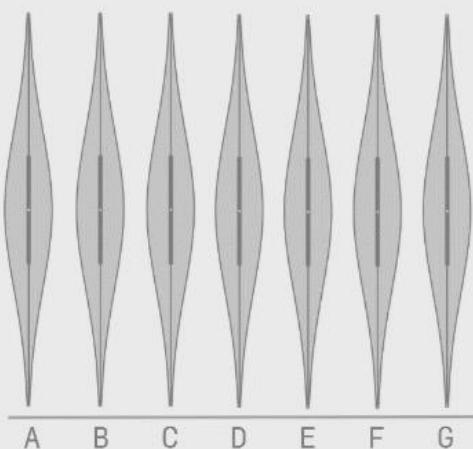


Symmetric (normal) vs skewed and uniform distributions



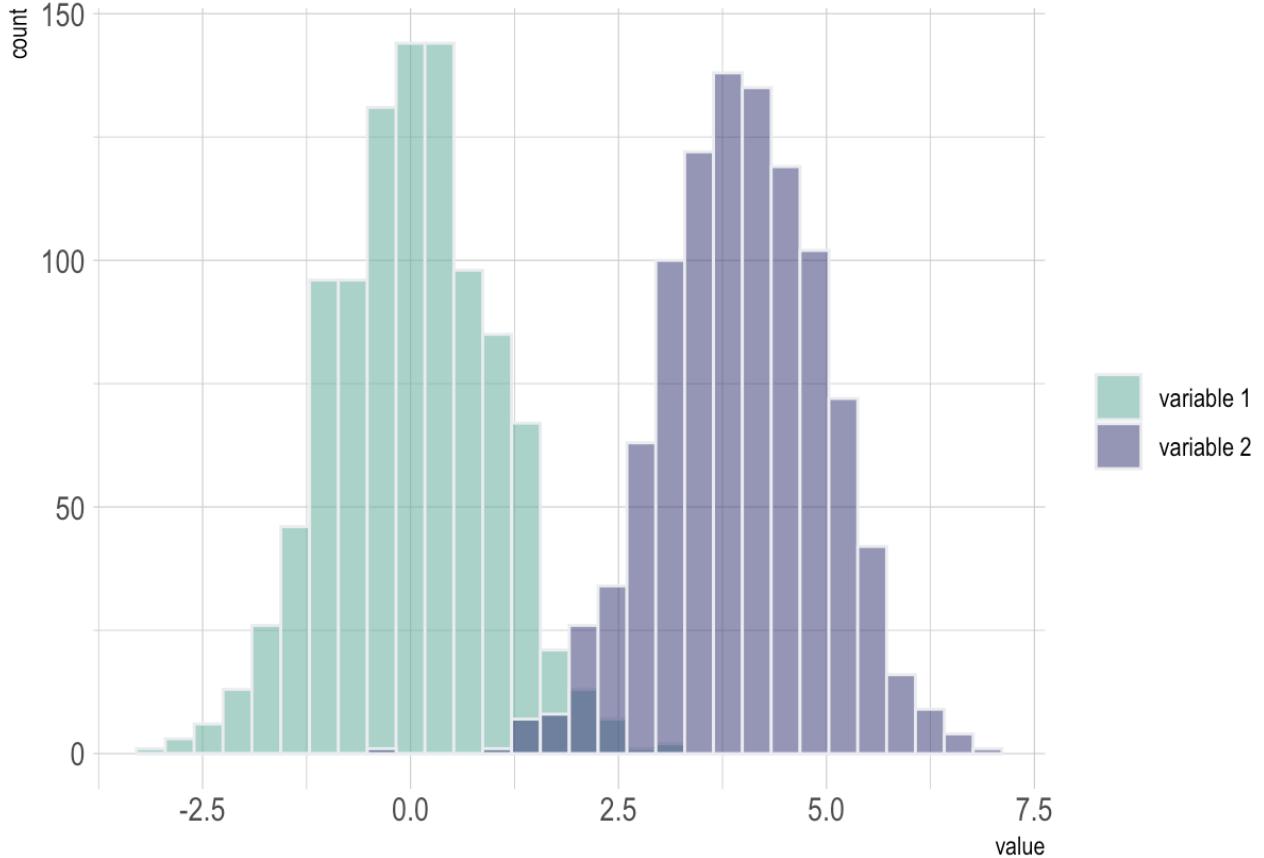
Unimodal vs bimodal distributions



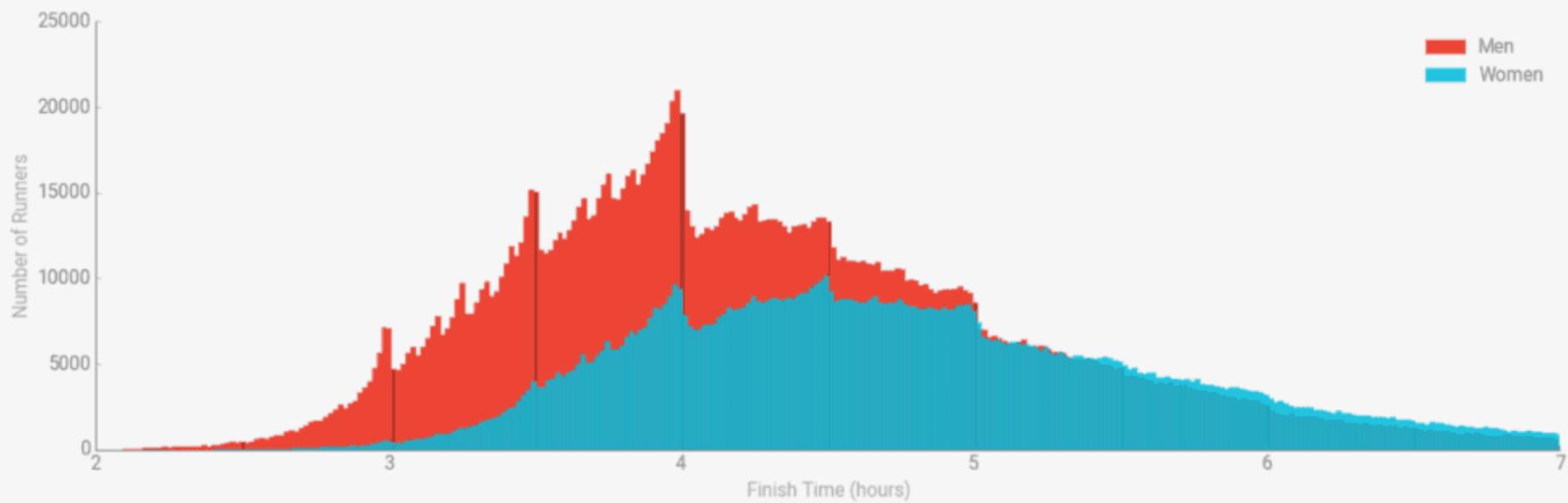
Raw Data**Box-plot of the Data****Violin-plot of the Data**



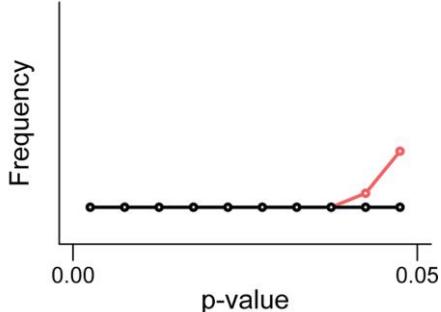
University
of Exeter



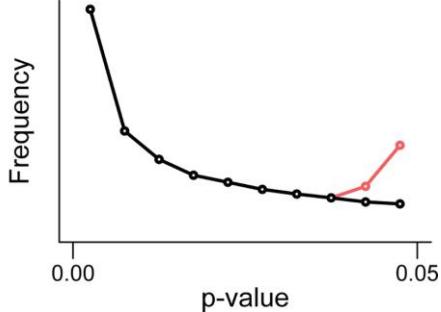
FINISH TIME DISTRIBUTIONS



A)



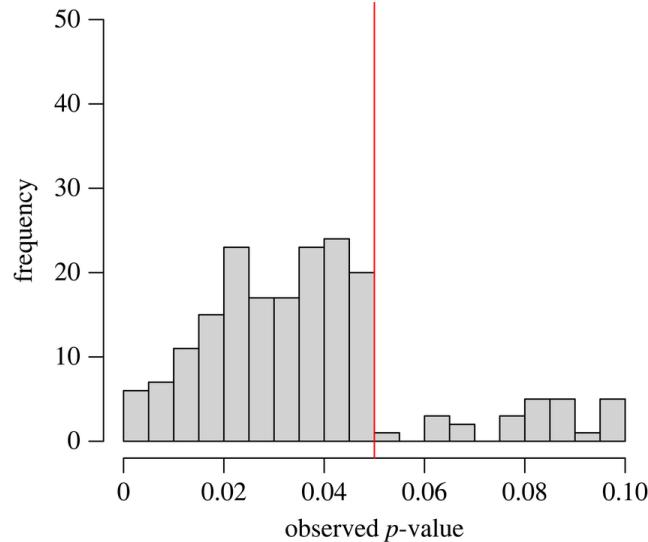
B)



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

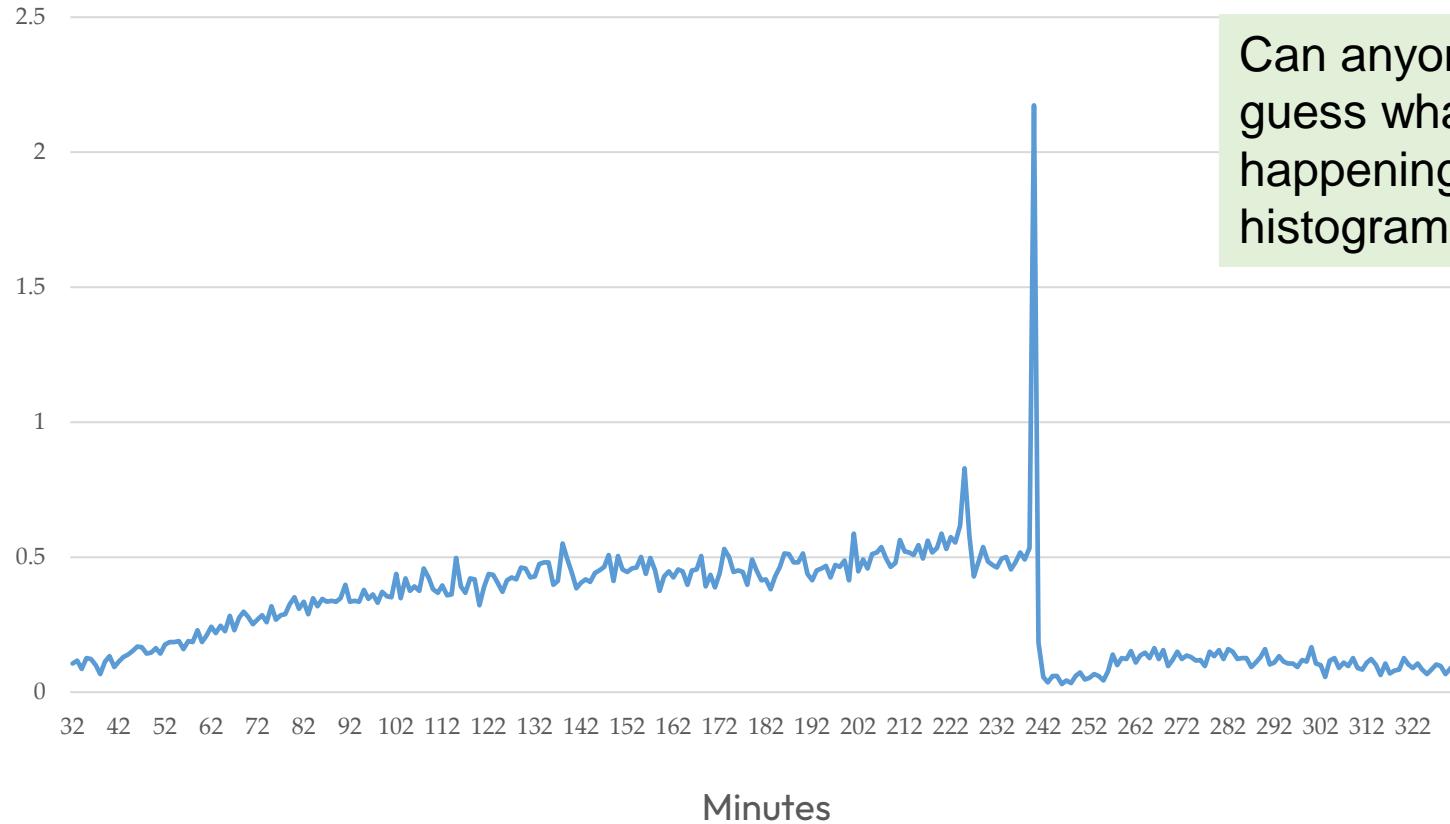
P-hacking can involve looking at many relationships, collecting or selecting data, or choosing different methods until non-significant results become significant.

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



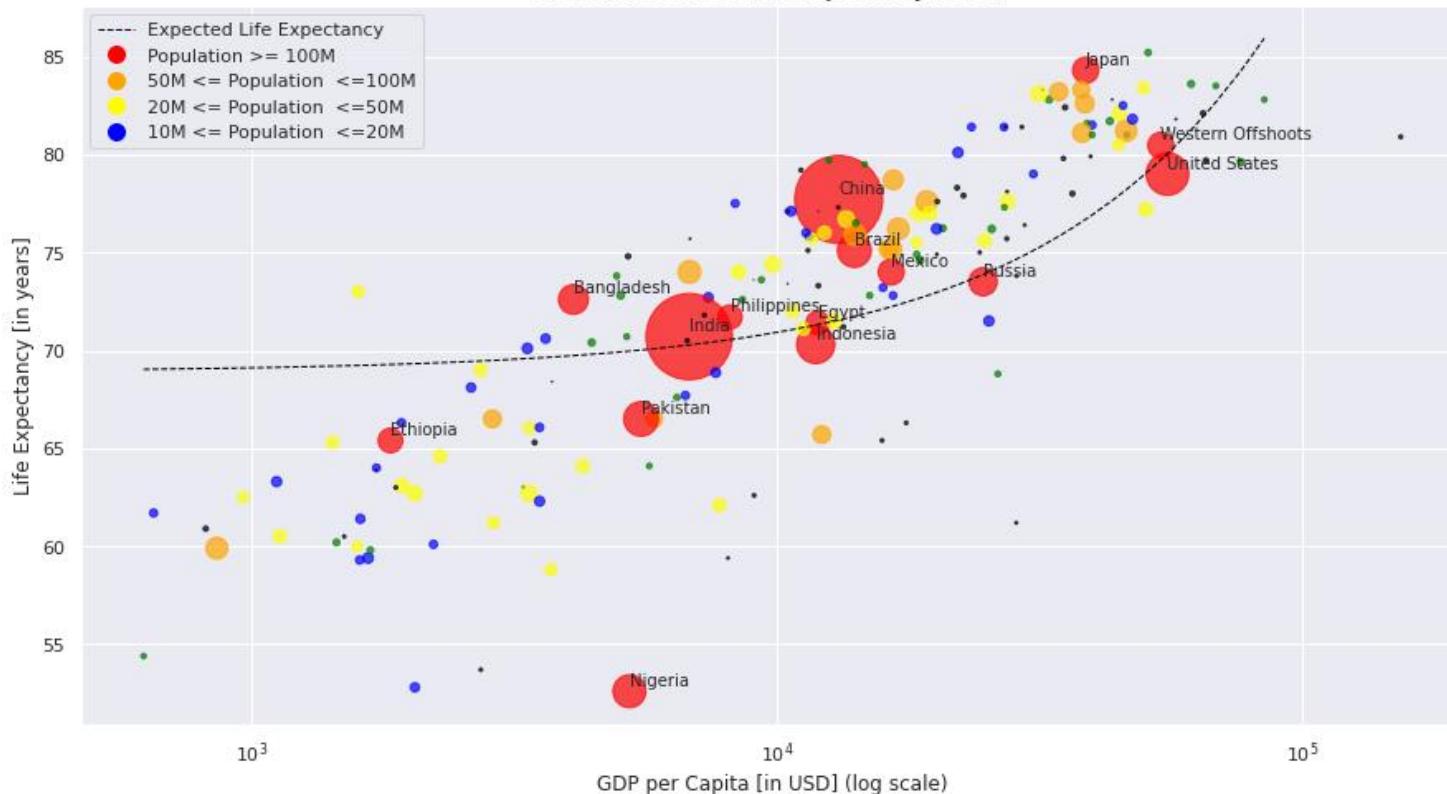
Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science: a narrative review of selected methodological issues in the field. *Royal Society Open Science*, 9(12), 220946.

Emergency Department length of stay



Can anyone
guess what is
happening in this
histogram?

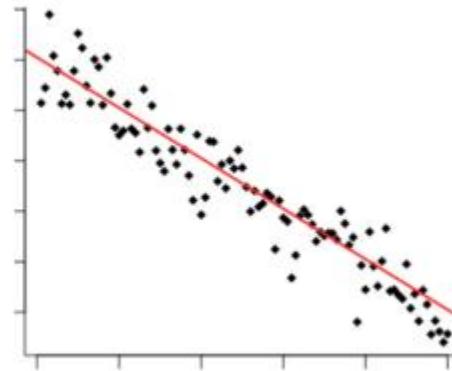
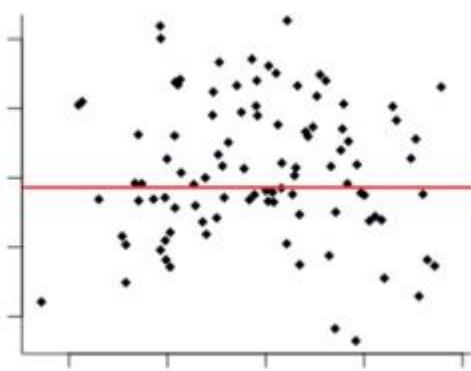
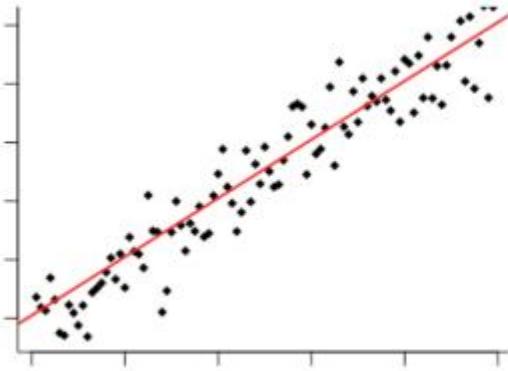
Worldwide GDP vs Life Expectancy (2018)



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

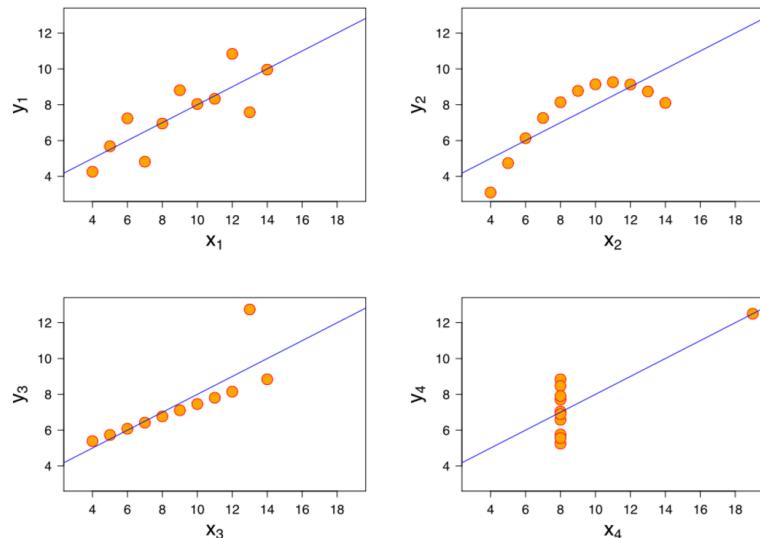
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	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)

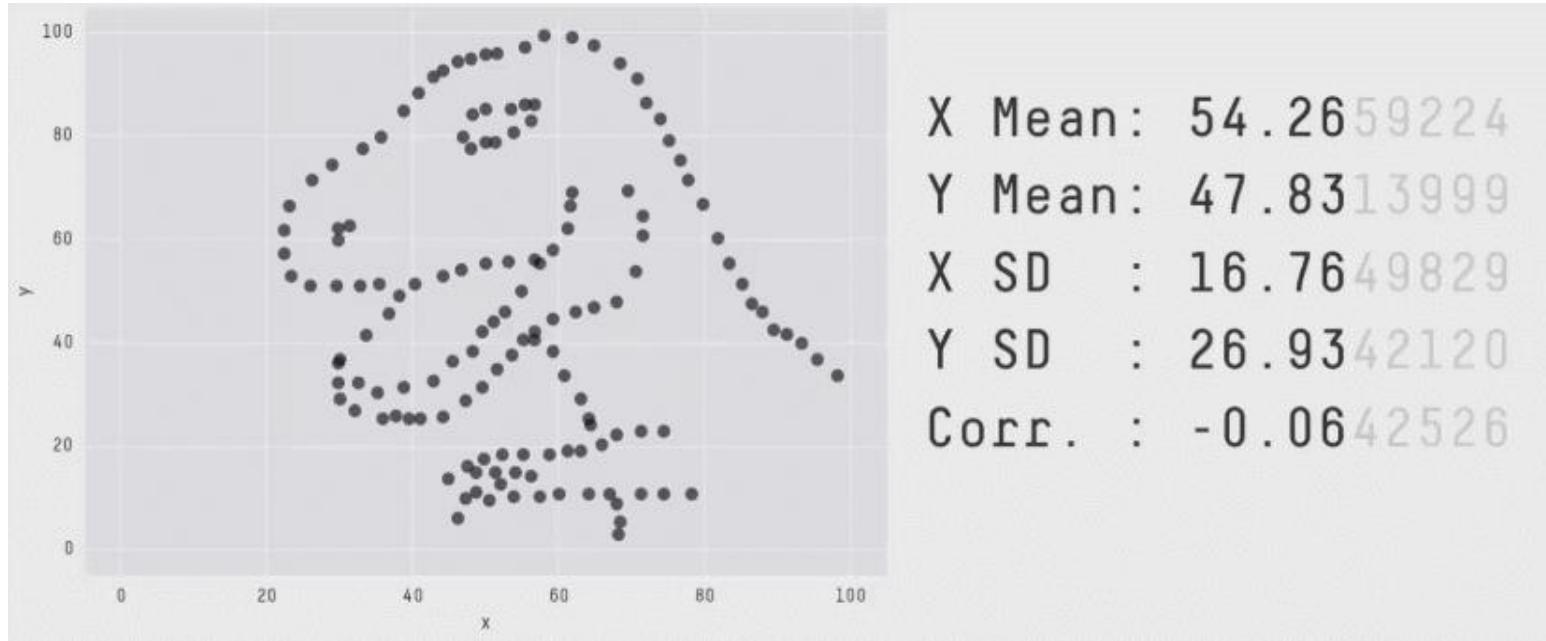
[Source](#)

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.

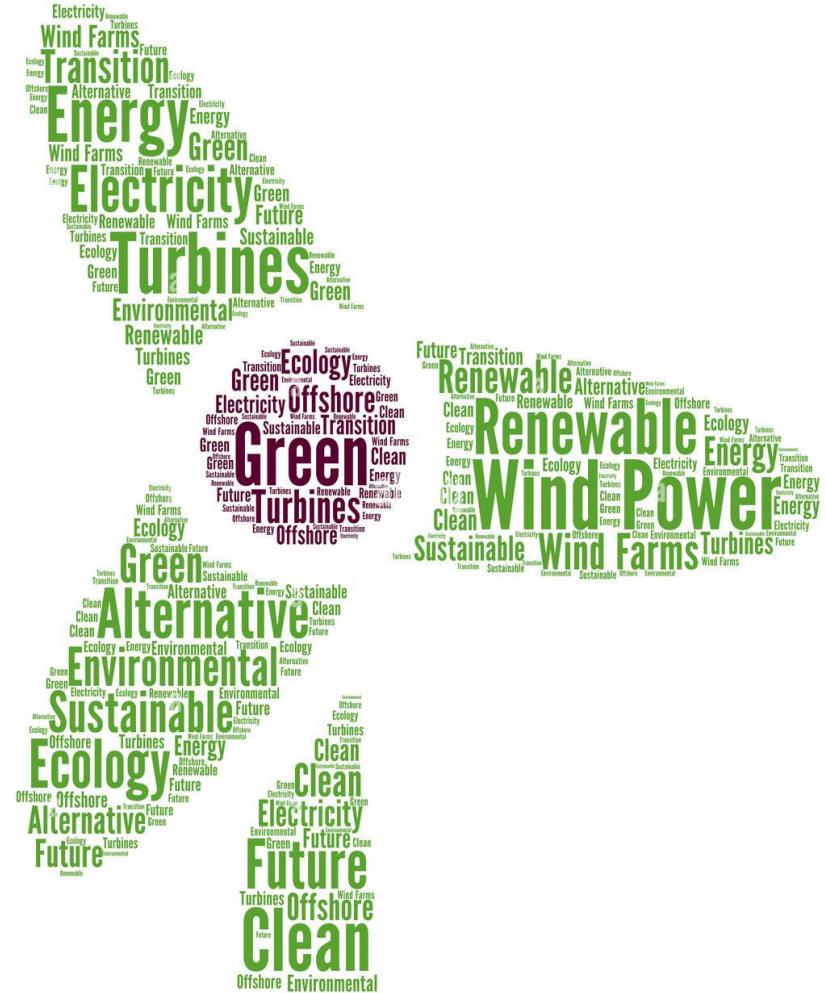


"Same Stats, Different Graphs: Generating Datasets With Varied Appearance and Identical Statistics Through Simulated Annealing," by J. Matejka and G. Fitzmaurice, *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (<https://doi.org/10.1145/3025453.3025912>). Copyright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

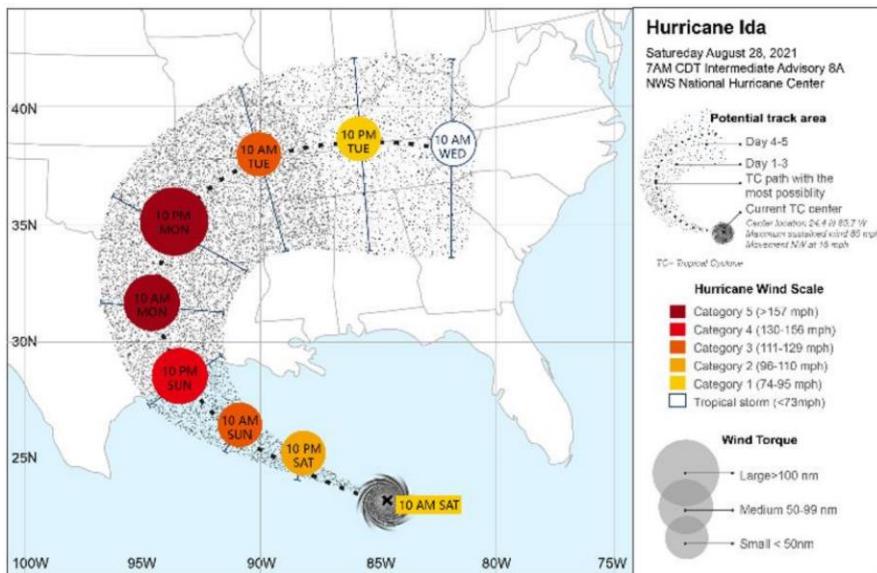
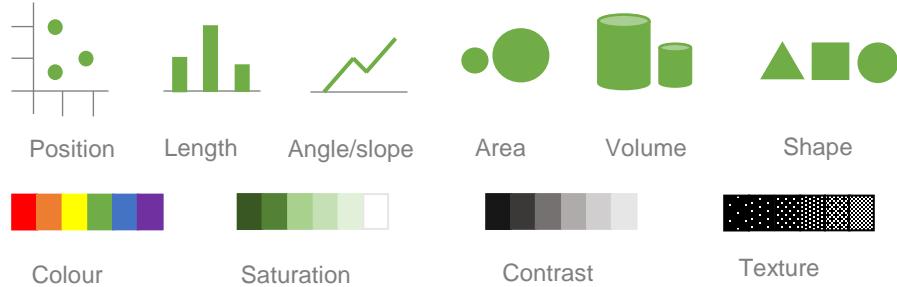


University
of Exeter

Visual vocabulary



Visually Encoding Data



Data storytelling allow us to use data sets to convey insights effectively.

Visual data encoding helps people to quickly understand data by enhancing **pattern recognition**, **reducing cognitive load** by making intuitive insights, and **preventing misinterpretation**.

The choice of encoding type can significantly impact how the audience perceives and interprets the data story.

Effective encoding captures attention, clarifies complexities, and emphasizes key messages.

Why do people hate Pie Charts?



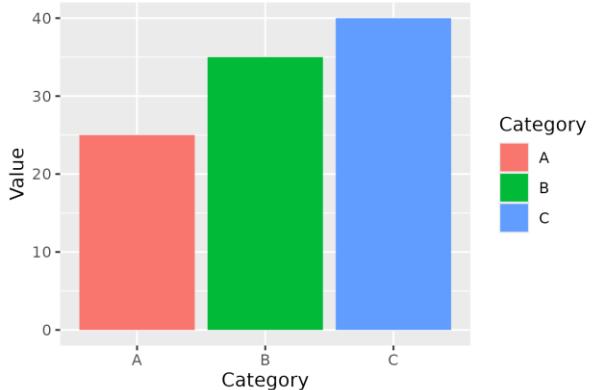
University
of Exeter

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

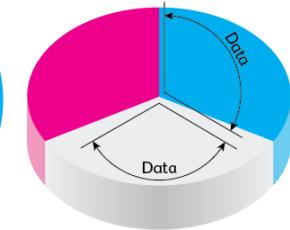
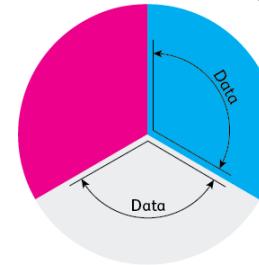
Pie Chart



Bar Chart



Angle



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

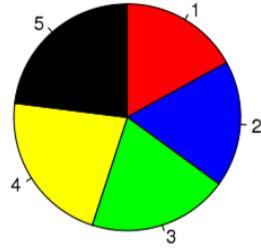
You may want to say it's encoded in the area, but if you create a pie chart by hand, what's the first thing you need to do?

Looking at Pie Charts....

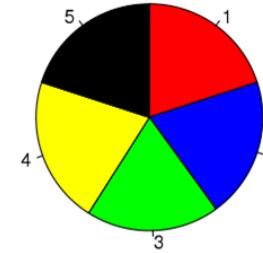


University
of Exeter

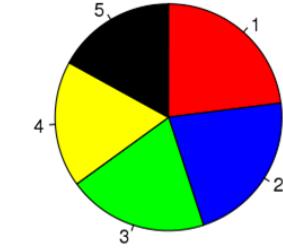
A



B



C

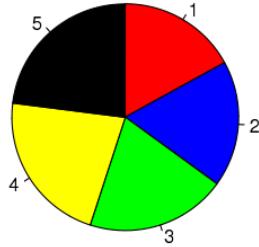


Looking at Pie Charts....

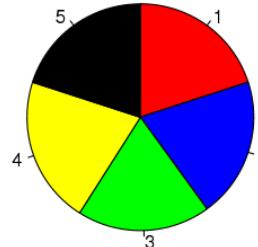


University
of Exeter

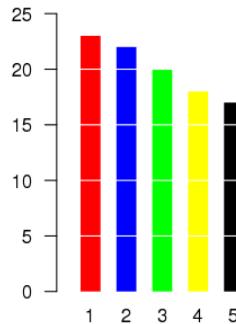
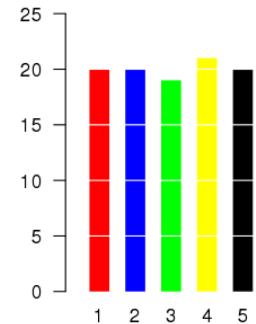
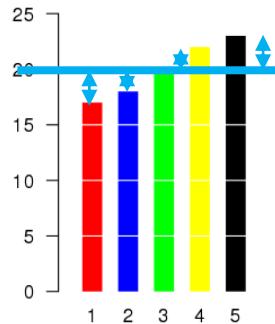
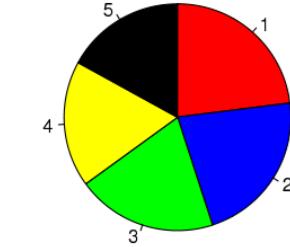
A



B



C

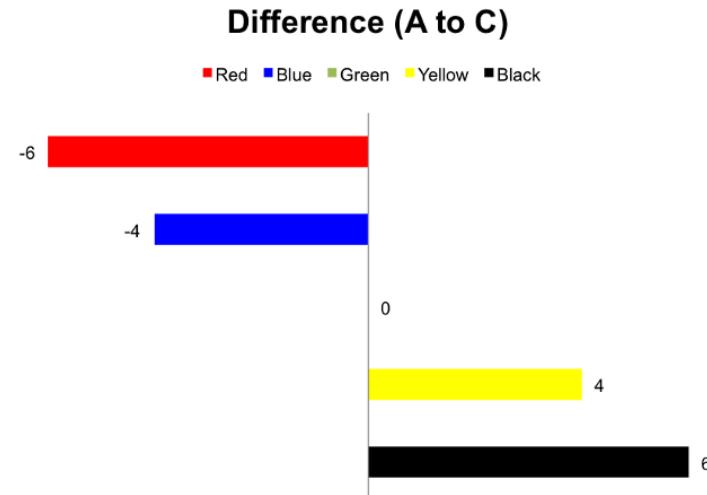


Source

Looking at Pie Charts....



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



Looking at Pie Charts....



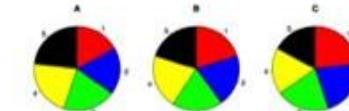
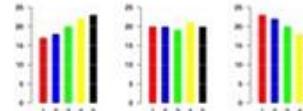
University
of Exeter

How important
is it to see
those small
differences?



Easy to determine
relative differences:

*"Black changed more than
yellow"*



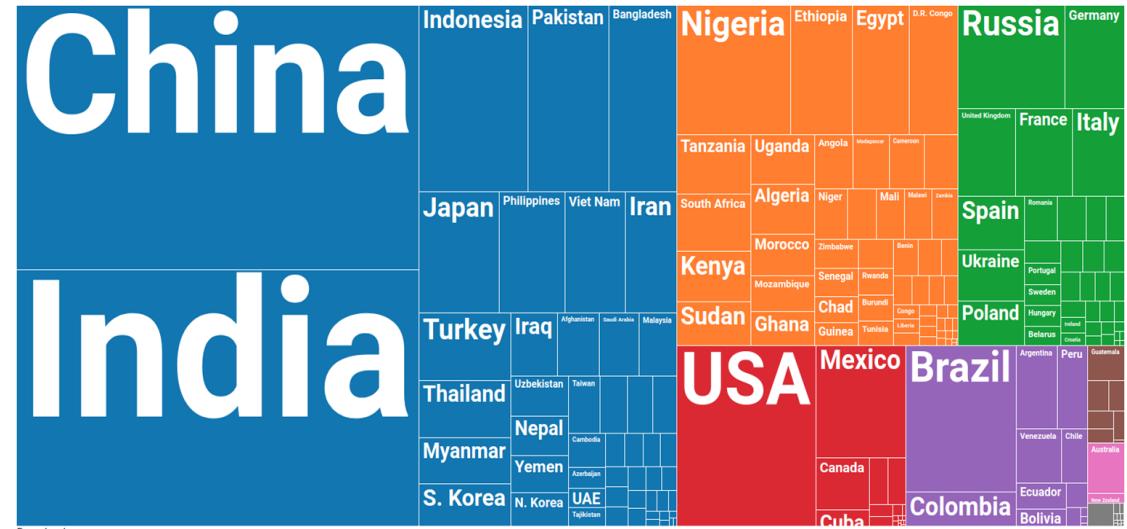
Easy to determine
absolute differences:

*"Neither yellow nor black
changed much"*

Proportions. Tree maps.

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

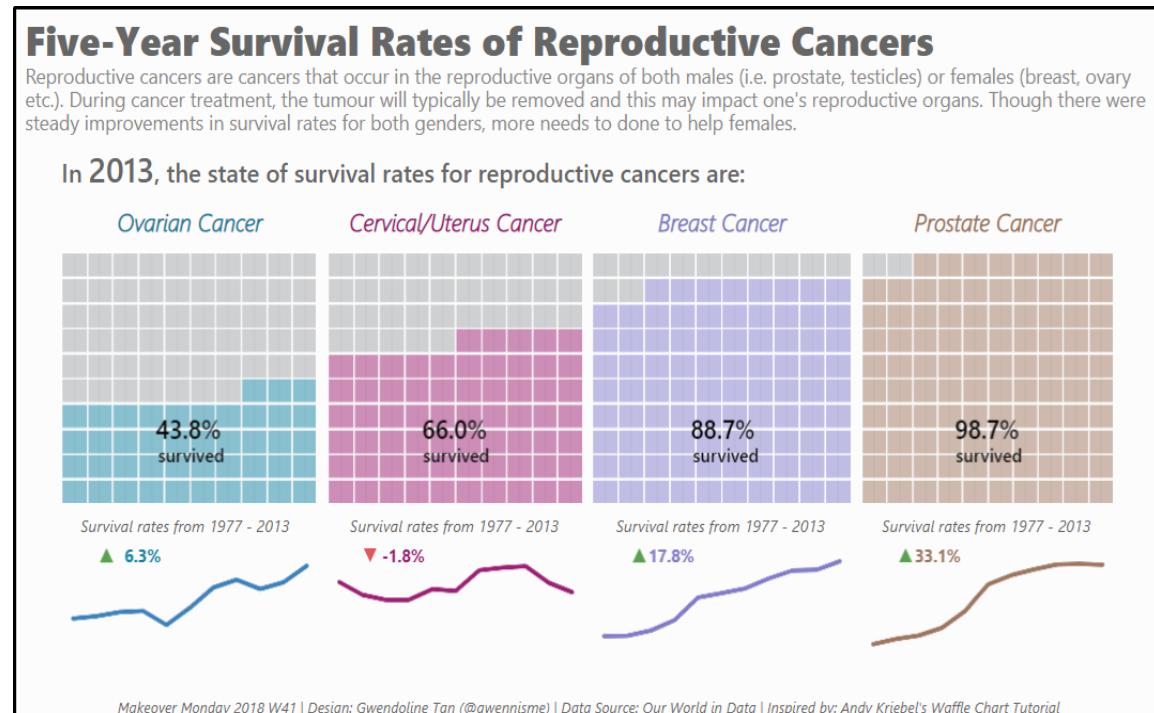
See this [interactive map](#) of population.



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.

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



What *is* 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

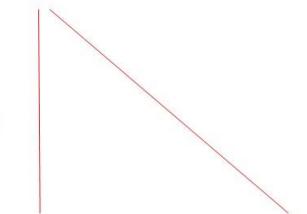
What did we test?

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**

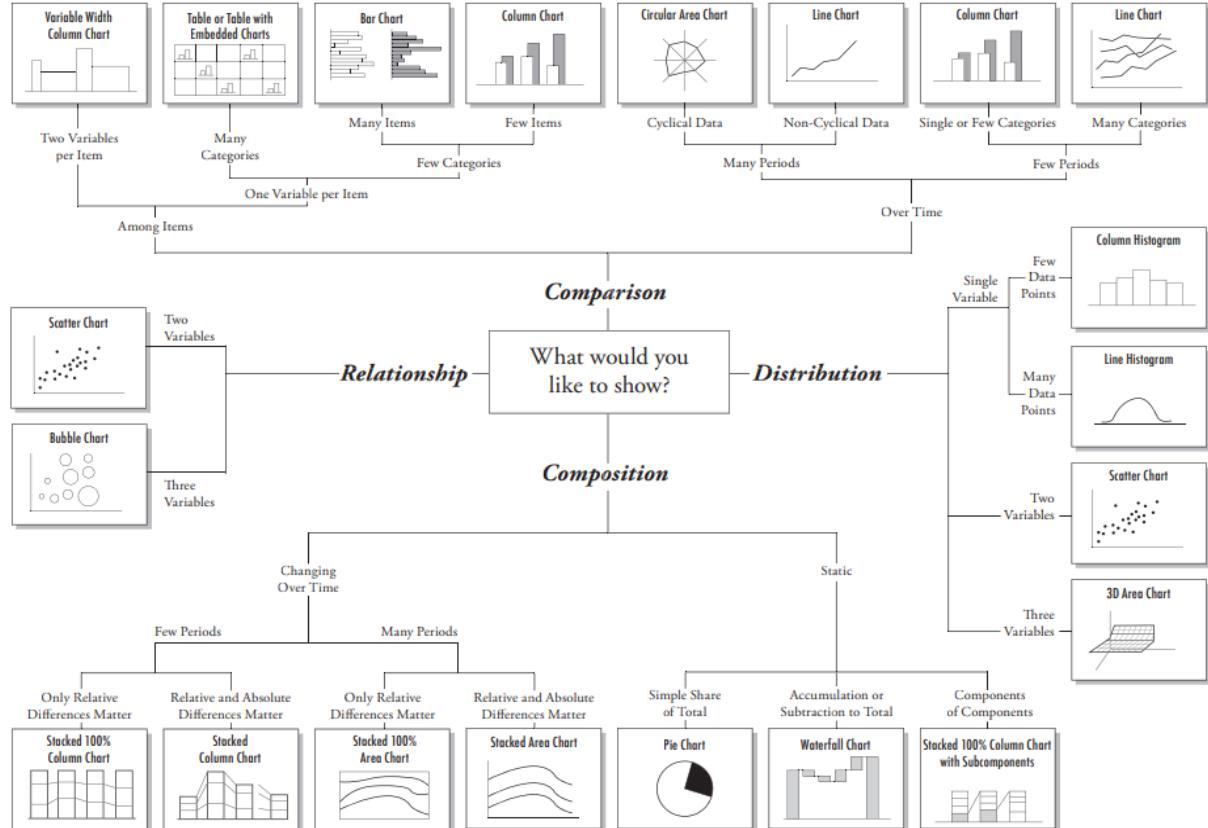
Stroke onset time was not determined when it potentially could have been



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



Chart Suggestions—A Thought-Starter





Deviation

Opposite variability (it's how a field varies from one point to another). The more it varies, the more it can also be a larger area. It's important to note that not all areas have the same amount of variation.

Example PT uses

Time series, movement, income and the economy

Correlation

Show the relationship between two or more variables. It's important to note that not all variables have the same correlation. Some are positive, some are negative, and some are neutral.

Example PT uses

Health, demography, budget tables, communications research

Ranking

Use where there's position in an ordered list. It's important to note that not all variables have the same ranking. Some are higher, some are lower, and some are in the middle.

Example PT uses

Wealth, demography, budget tables, communications research

Distribution

Show values in a dataset and how often they occur. This can be seen as a histogram or a bell curve. It's important to note that not all distributions are normal.

Example PT uses

Demography, population

Change over Time

Show estimates to changing trends. This can be seen as a line graph or a scatter plot. It's important to note that not all trends are linear.

Example PT uses

Demography, environment, new data series, seasonal changes in a market

Magnitude

Show size comparisons. This can be seen as a bar chart or a pie chart. It's important to note that not all magnitudes are equal.

Example PT uses

Communication, market capitalisation, volumes in general

Part-to-whole

Show how a single entity can be broken down into smaller parts. It's important to note that the reader's interest is often in the size of the whole, not the size of the magnitude type chart instead.

Example PT uses

Face, budget, company structures, national election results

Spatial

Show from location, volume or intensity of data. It's important to note that not all locations are created equal. There might be more points in one location than others, or more points in one geographical location than others.

Example PT uses

Demography, natural resources, locations, industrial distribution, health, education, areas, variation in election results

Flow

Show the relative volume or intensity of data moving from one place to another. It's important to note that not all sequences or geographical locations are created equal.

Example PT uses

Demography, trade, migration, flows between informants, relationship graphs

Giving bar

A simple standard bar chart showing both regular and irregular negative values.

Scatterplot

The easiest way to show the relationship between two continuous variables. It's important to note that not all scatterplots have their own axes.

Ordered bar

Standard bar chart where the values must make sense when sorted. It's important to note that not all bars have their own axes.

Histogram

The easiest way to show a distribution. It's important to note that not all histograms have their own axes.

Line

The easiest way to show a change in data over time. It's important to note that not all lines have their own axes.

Bar

The easiest way to show a change in data over time. It's important to note that not all bars have their own axes.

Stacked column/bar

A simple way of showing multiple categories but can be hard to read if there are more than five categories.

Basic bubble chart

The standard approach for bubbles. It's important to note that not all bubbles have their own axes.

Sankey

Show change in flow. It's important to note that not all flows have their own axes.

Giving stacked bar

Perfect for presenting survey results which have been broken down by city and gender.

Dot plot

A good way of showing the difference between two or more things.

Connected scatterplot

Usually used to show how the relationship between two variables has changed over time.

Ordered proportional symbol

Use when there are big differences between the individual values in a dataset.

Dot skip plot

Good for showing individual values in a dataset when the data is not as important.

Barcode plot

Like dot plots, good for showing different individual values in a dataset.

Calipers

Column well for showing change over time. It's important to note that not all calipers have their own axes.

Bar

Show above. Good when the data are not as important and the categories have long category names.

Matrices

A great way of showing the size and proportion of data at once. It's important to note that not all matrices have their own axes.

Name

Shows a single value representing something. It's important to note that not all names have their own axes.

Bubble

A bubble chart. It's important to note that not all bubbles have their own axes.

Dot strip plot

Shows data in order on a strip set against a background of many other things.

Dot plot

A good way of showing the difference between two or more things.

Line

Good for showing change over time. It's important to note that not all lines have their own axes.

Calipers

As per standard, column well for showing change over time. It's important to note that not all calipers have their own axes.

Paired bar

As per standard, column well for showing change over time. It's important to note that not all paired bars have their own axes.

Pie

A common way of showing part of a whole. It's important to note that not all pie charts have their own axes.

Dot

Similar to a dot plot but here the centre can't be seen. It's important to note that not all dots have their own axes.

Bar chart

A chart of bars. It's important to note that not all bar charts have their own axes.

XY heatmap

A good way of showing the pattern between two variables. It's important to note that not all XY heatmaps have their own axes.

Slope

Perfect for showing how something has changed over time or space.

Loftmap

Loftmaps draw more complex shapes than standard bar charts. It's important to note that not all loftmaps have their own axes.

Bar

Effective for drawing changing variables. It's important to note that not all bars have their own axes.

Violin plot

Similar to a box plot but more complex. It's important to note that not all violin plots have their own axes.

Area chart

Use with area – these charts can change but the components can be compared.

Hexbin

A great way of showing the size and properties of data at once. It's important to note that not all hexbins have their own axes.

Proportional control

Use when there are big differences between values and seeing the overall picture is not too important.

Violin

Violin plot. It's important to note that not all violin plots have their own axes.

Candlesticks

Usually focused on financial markets. It's important to note that not all candlesticks have their own axes.

Dot chart

Use with area – these charts can change but the components can be compared.

Dot chart

Use with area – these charts can change but the components can be compared.

Fan chart

Use with area – these charts can change but the components can be compared.

Dot chart

Use with area – these charts can change but the components can be compared.

Timeline

Use when there are big differences between values and seeing the overall picture is not too important.

Network

A way of turning relationships into a network. It's important to note that not all networks have their own axes.

Scalped cartogram

Stretching each area to an area that's proportional to its value. It's important to note that not all scalped cartograms have their own axes.

Bar density

Used to show the location of individual individuals. It's important to note that not all bar densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Dot density

Used to show the location of individual individuals. It's important to note that not all dot densities have their own axes.

Frequency polygons

For displaying multiple distributions of data. It's important to note that not all frequency polygons have their own axes.

Boxscore

Boxscore. It's important to note that not all boxscores have their own axes.

Connected scatterplot

A good way of showing the difference between two variables. It's important to note that not all connected scatterplots have their own axes.

Calendar heatmap

A good way of showing the difference between two variables. It's important to note that not all calendar heatmaps have their own axes.

Probability timeline

Great when data is categorical. It's important to note that not all probability timelines have their own axes.

Circle timeline

Good for showing the difference between two variables. It's important to note that not all circle timelines have their own axes.

Vertical timeline

Presents time on the Y axis. It's important to note that not all vertical timelines have their own axes.

Timeline

Another way of showing the difference between two variables. It's important to note that not all timelines have their own axes.

Timeline

Another way of showing the difference between two variables. It's important to note that not all timelines have their own axes.

Salogram

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Stemograph

A type of stem chart. It's important to note that not all stemographs have their own axes.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Grouped bar

Another after-effect. Another after-effect for the circle timeline for when there are big sections in the data.

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to help you think about what's most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

Source

© 2017 Alex Southgate. Used under a Creative Commons Attribution Non-Commercial-ShareAlike license. All rights reserved.

QR code

ft.com/vocabulary



© FT.com

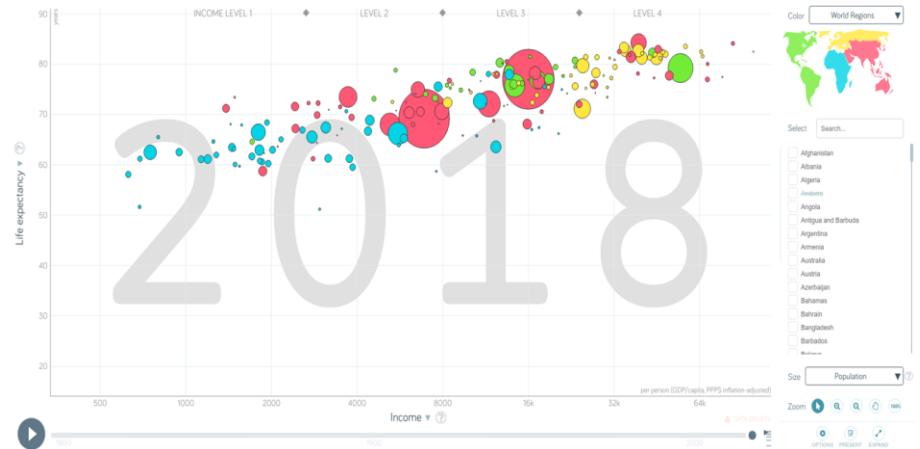
There are a huge range of different ways of showing data. The Financial Times created this chart of a Visual Vocabulary to help give names to the different methods.

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.

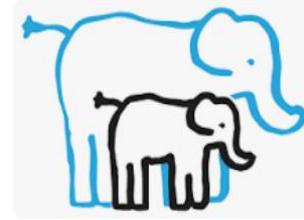
Gapminder has continued on without Hans' input since his passing and continued to help put world development and human experience data in context.





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

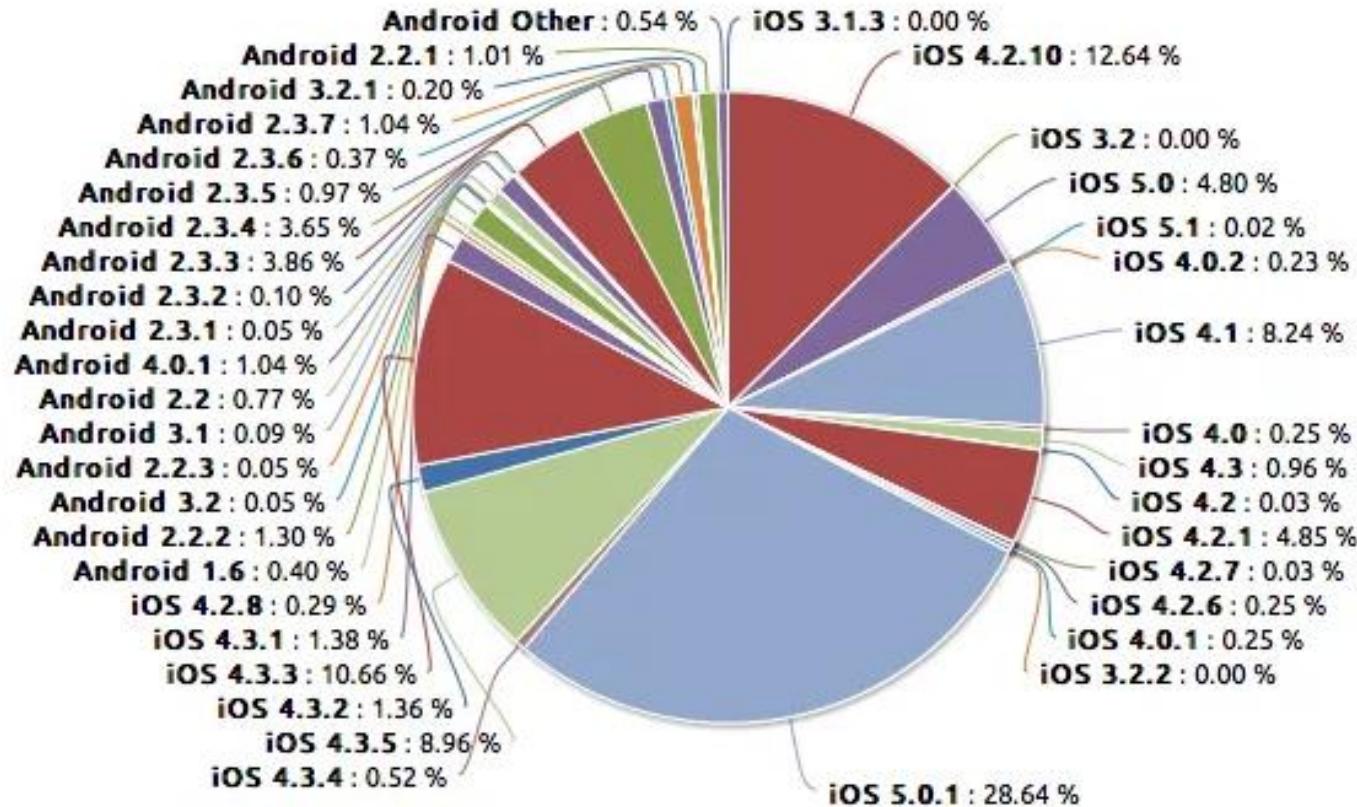
- 1. Clarity and Simplicity:** Keep the visualization simple and easy to understand. Avoid clutter, excessive decorations, and non-essential elements. The audience should be able to grasp the main message quickly.
- 2. Accuracy:** Ensure that the data presented is accurate and that the visualization 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.



University
of Exeter

Chart Junk

Crashes by OS Version Normalized (12/1 - 12/15)





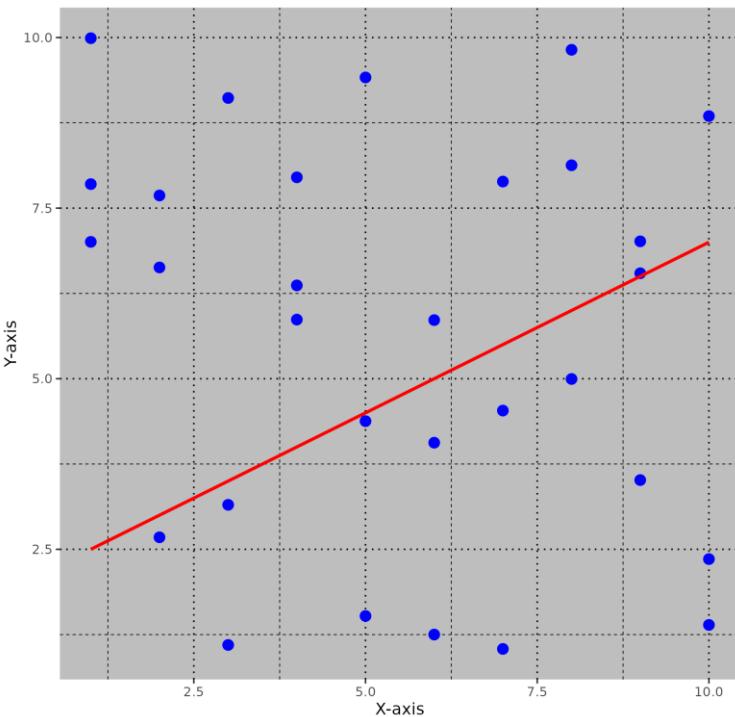
Ink to Data Ratio

- 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.

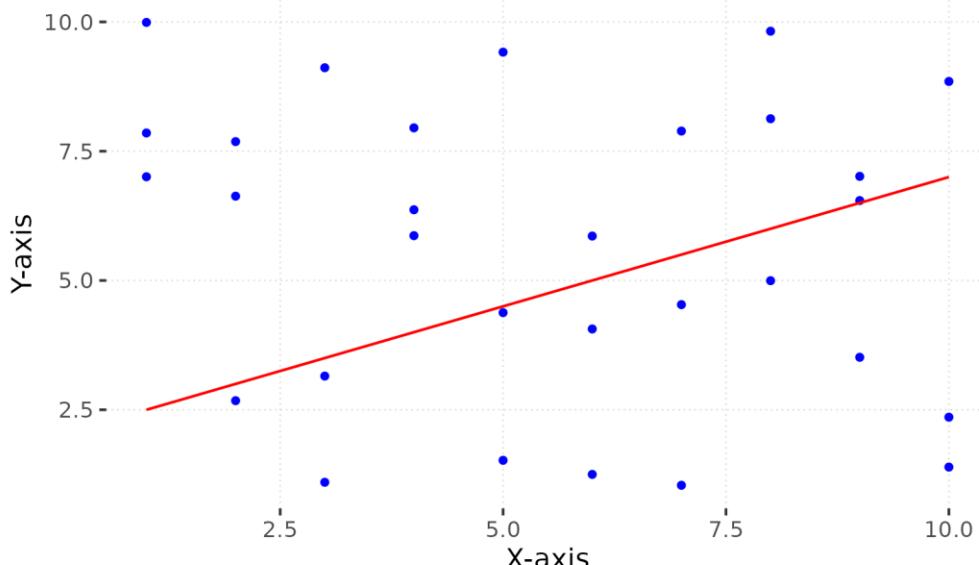
Principles of Data Ink

Above all else show data.

Scatterplot and Line Plot with Busy Grid



Scatterplot and Line Plot with Clean Grid



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

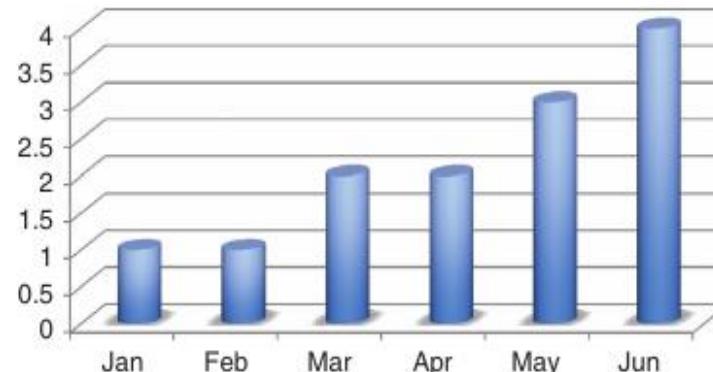
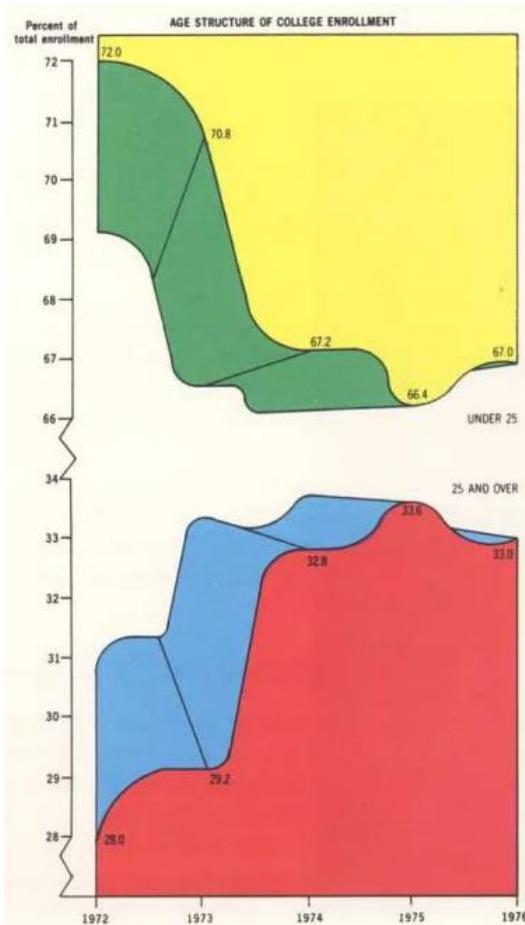


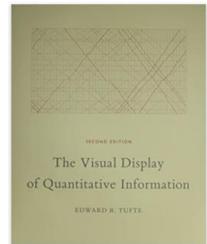
FIGURE 2.25 3D column chart



*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).*

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

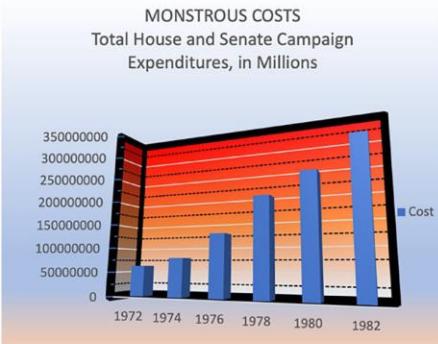
Tufte, E. R. (2001).
The visual display of quantitative information, Graphics press Cheshire, CT.



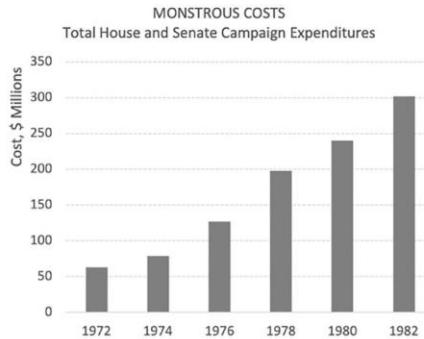
Tufte (1983, p.118) says, "This may well be the worst graphic ever to find its way into print."



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



A “cluttered” visualization (top),



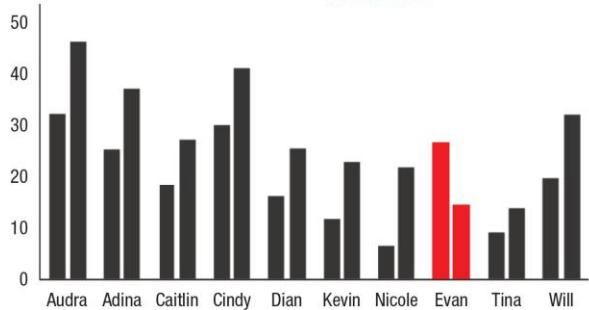
a minimalist “decluttered” version (middle),

and a version that incorporates pictorial embellishment (bottom).



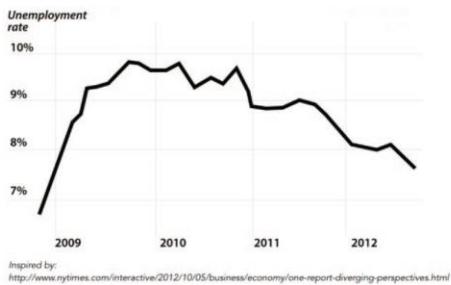
The graph at the bottom was created by Nigel Holmes for *TIME Magazine* and was reprinted in his [1984](#) book, *Designer's Guide to Creating Charts & Diagrams*.

One Student Got Worse

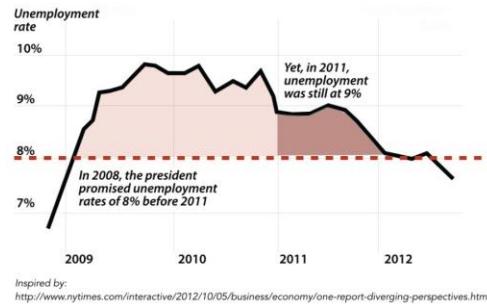


Unemployment is higher than stated goals

In 2008, the president promised unemployment rates under 8% before 2011. Yet, in 2011, unemployment was still at 9%

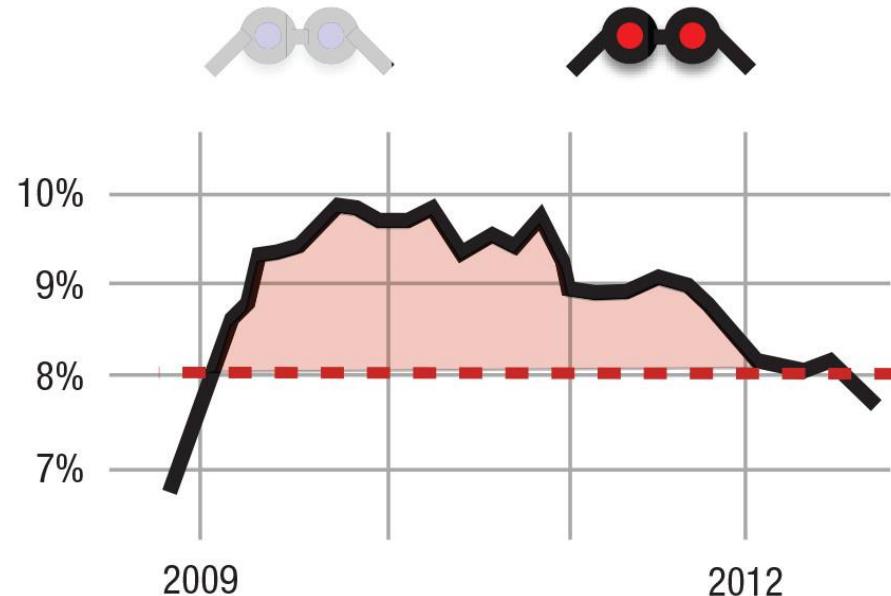
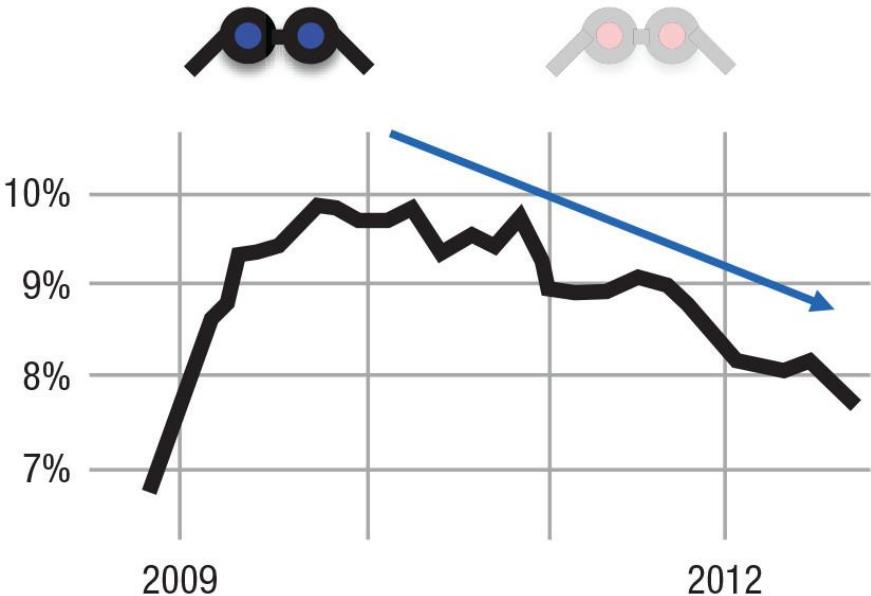


Unemployment is higher than stated goals



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

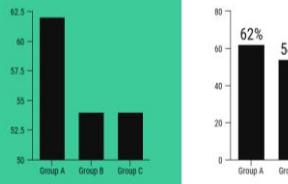


An example of emphasizing different perspectives in a single data set (inspired by [Bostock et al., 2012](#)). One data set can be seen with dramatically different perspectives, depending on which patterns an observer does and does not extract.

Using graphs to mislead

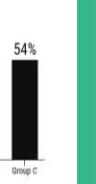
1 ➤ OMITTING THE BASELINE

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.

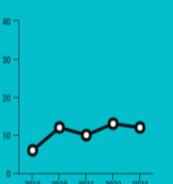


ACCURATE 😊

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

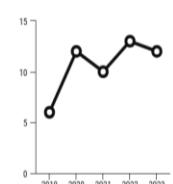
2 ➤ MANIPULATING THE Y-AXIS

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.

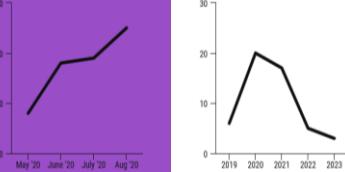


ACCURATE 😊

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

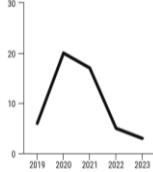
3 ➤ CHERRY PICKING DATA

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.

4 ➤ 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.
- This graph can confuse and mislead readers, who expect dark to represent a higher population density.

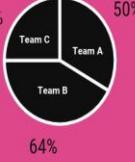


ACCURATE 😊

- 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.

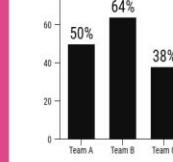
5 ➤ USING THE WRONG GRAPH

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.

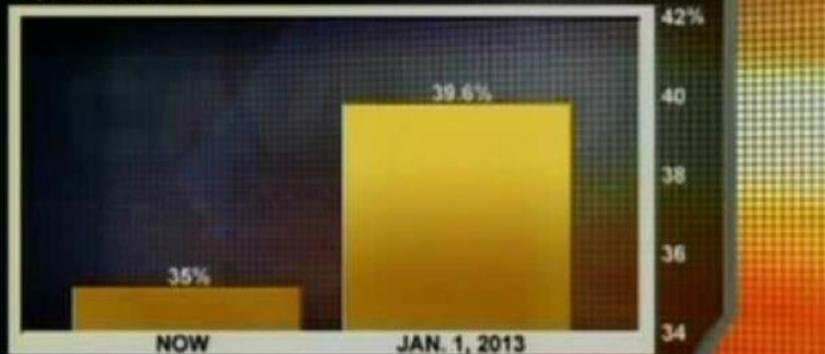


ACCURATE 😊

- 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

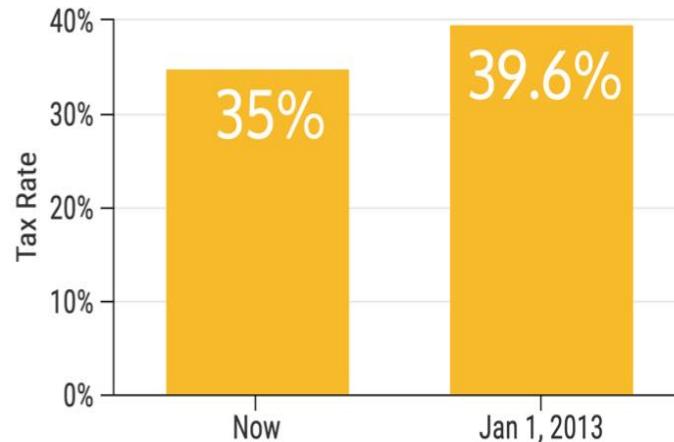
WITH THE JUSTICE DEPARTMENT AND ACQUIRES FULL T

DOW 13008.68 □ 64.33

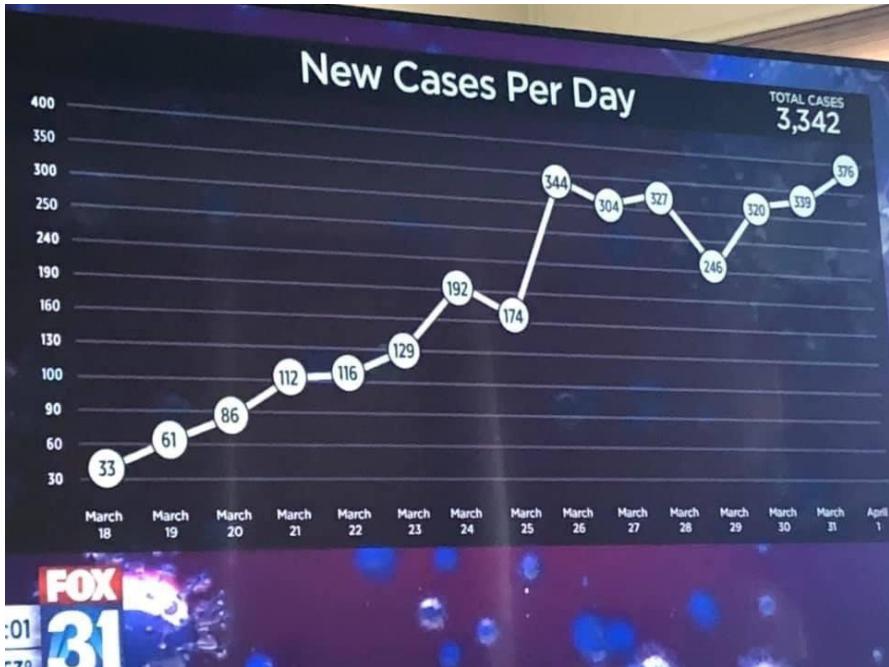
S&P 1379.32 □ 5.98

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 seems to be completely arbitrary.

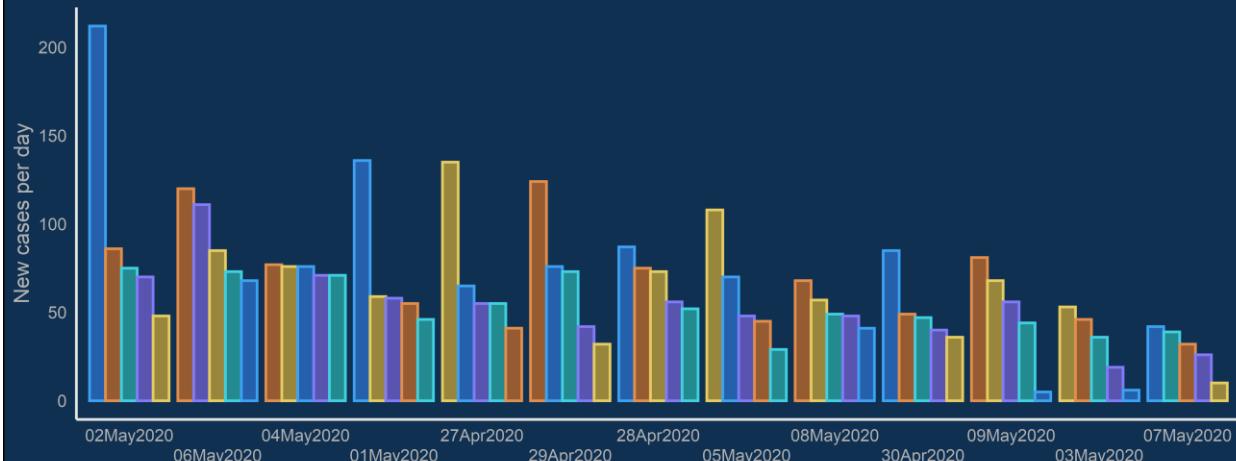
[How to make that crazy Fox News y axis chart with ggplot2 and scales \(freerangestats.info\)](https://freerangestats.info/)

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
\(freerangestats.info\)](http://freerangestats.info)

Absolute Precision Ranking for Seeing a Single Ratio

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

Position: 1 (1), 7 (7)

Length: 1 (1), 7 (7)

Area: 1 (1), 7 (7)

Angle: 1 (1), 7 (7)

Intensity: 1 (1), 7 (7)

Highest ↑ Lowest ↓

Common Illusions That Distort Data

Caveats for the visual encoding in each row

99 99
98 98
a b a b

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

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

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

Sure, looks like a ~1:7 area ratio.

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

Intensity values can look different depending their backgrounds.
Do not plot intensities on intensities.

Vision Is Powerful for Global Statistics

For each visualization, statistics are available quickly

Dot Plot: Max height, Mean height, Min height

Stacked Bar: Min, Mean length of dark bars, Max

Bubble Map: Mean Area, Min, Max

Slope Graph: Max, Mean Angle, Min

Heat Map: Mean Intensity, Min, Max

Vision Is Sluggish for Comparisons

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

So guide viewers to the right comparisons

a b c d e f

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

a b c d e f

“a, c, & e have increased”

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

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

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



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?

?