University of Exeter

Data Visualisation

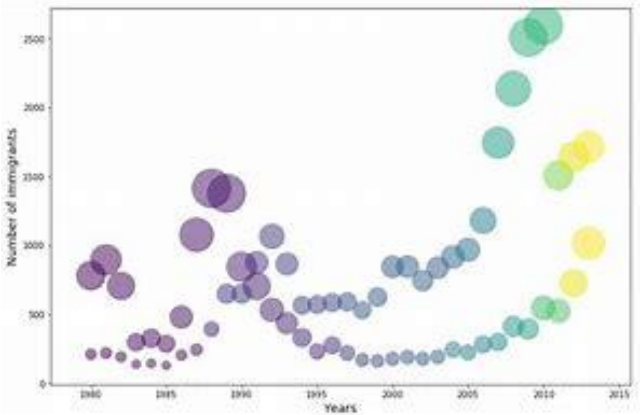
Week 03-BEM2031

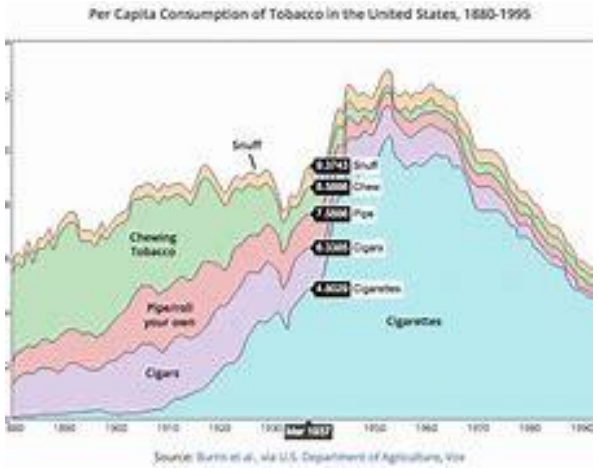
Term2: 2023/24

University of Exeter

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

University of Exeter

Histograms

Statistics

(summary statistics)

Descriptive Statistics

Inferential Statistics

Statistics

(summary statistics)

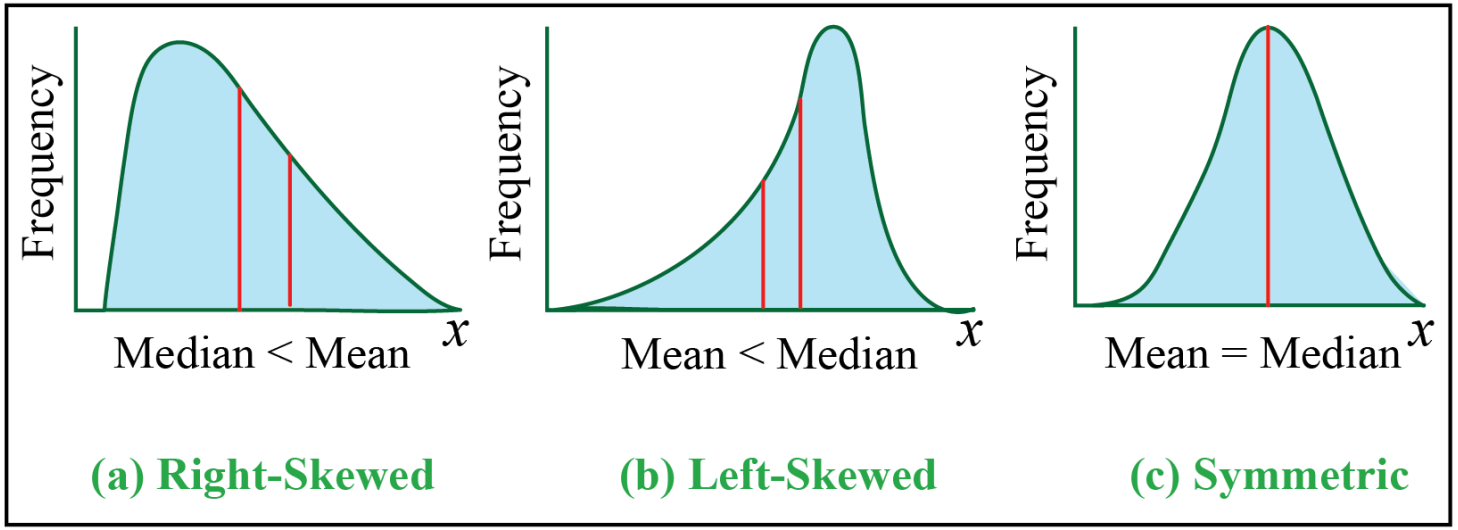
Descriptive Statistics		Inferential Statistics
Central tendency	• Mean • Median • Mode	
Dispersion / variability	• Range • Interquartile range • Variance • Standard deviation	
Skewness	• Symmetric • Left • Right	

Statistics

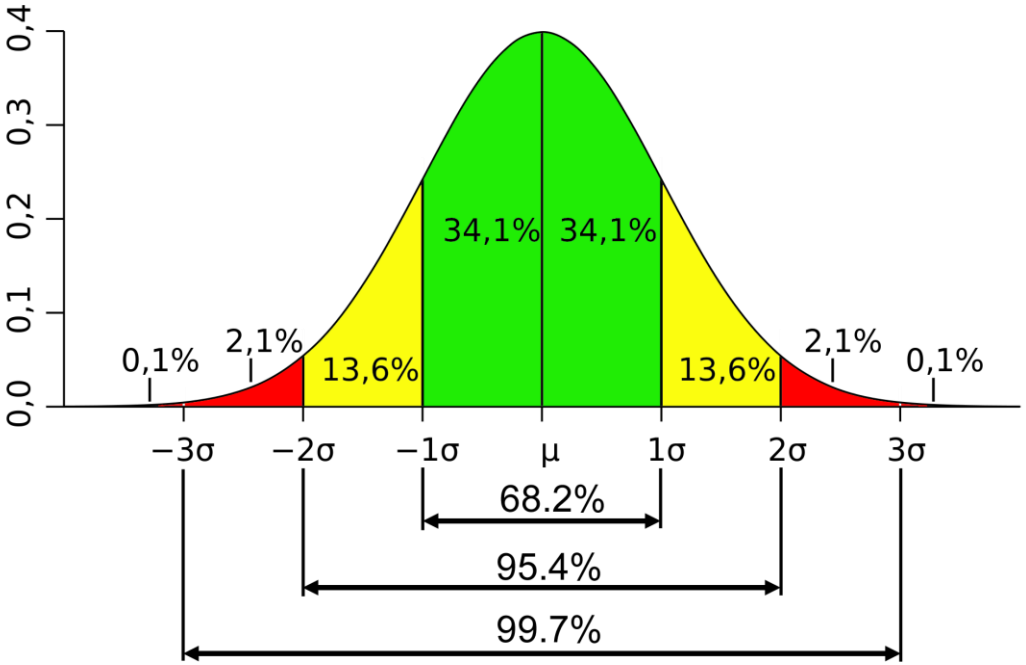
(summary statistics)

Descriptive Statistics		Inferential Statistics
Central tendency	• Mean • Median • Mode	Estimation parameters
Dispersion / variability	• Range • Interquartile range • Variance • Standard deviation	Hypotheses testing
Skewness	• Symmetric • Left • Right	Tests of difference and similarity

Statistics

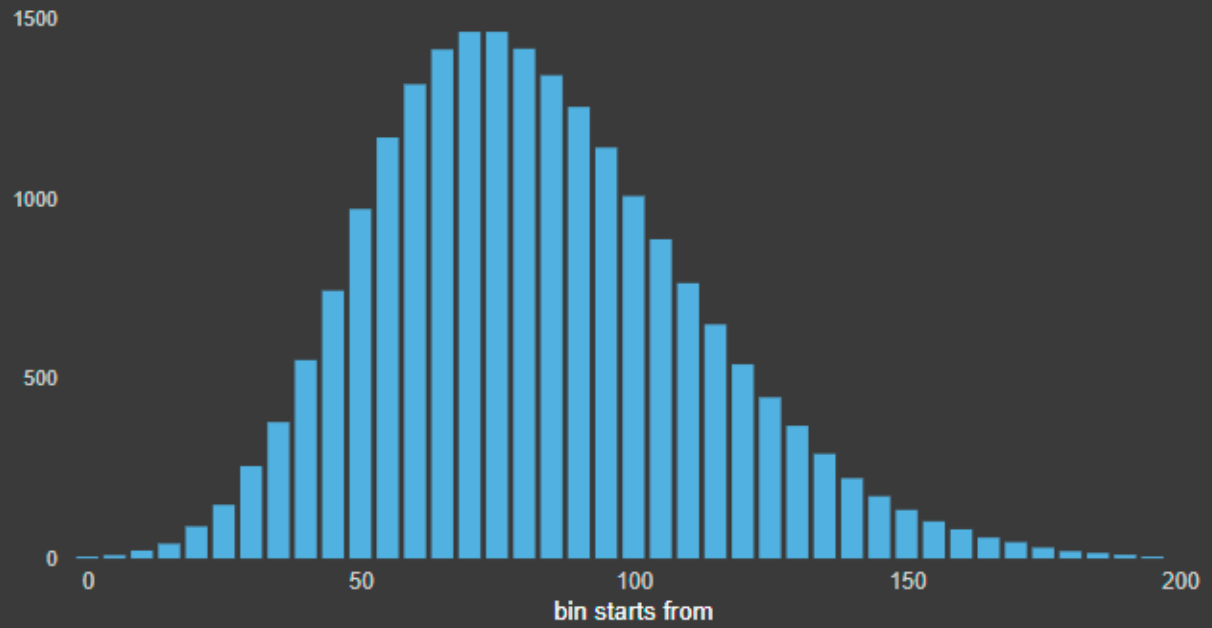

Frequency
Median < Mean
x
(a) Right-Skewed
Frequency
Mean < Median
x
(b) Left-Skewed
Frequency
Mean = Median
x
(c) Symmetric

Statistics



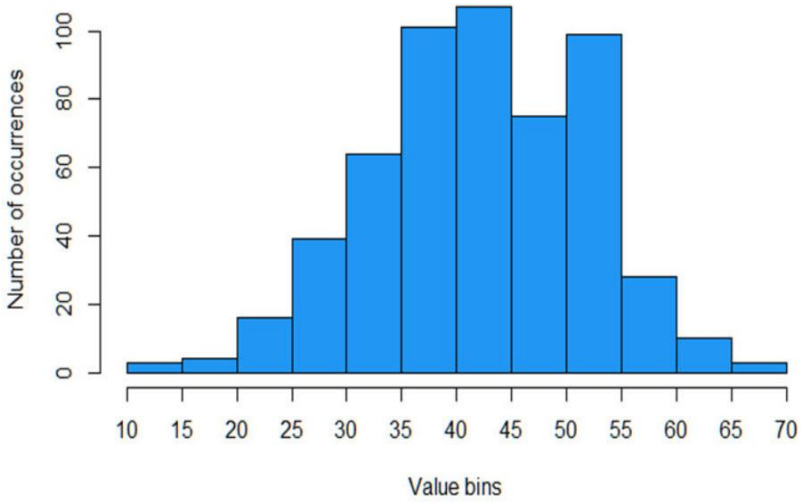
Variance: the average squared differences from the mean

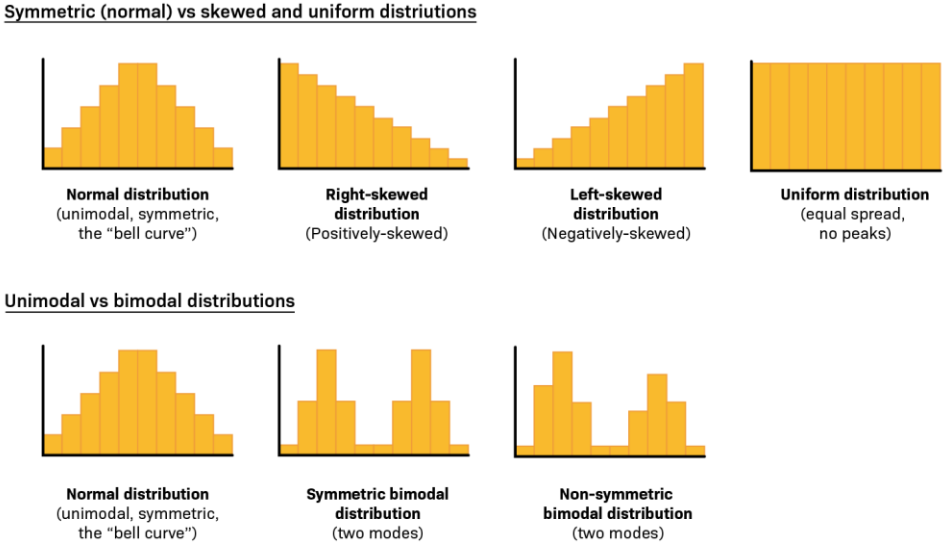
Standard deviation: the square root of the variance

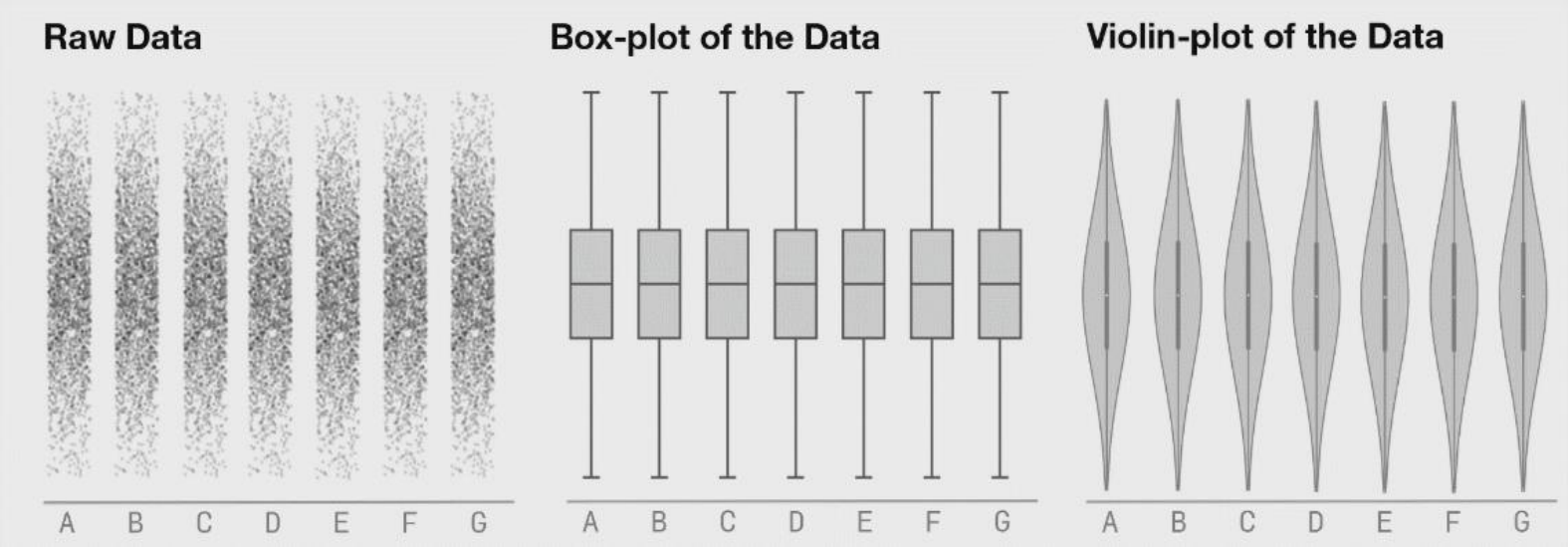


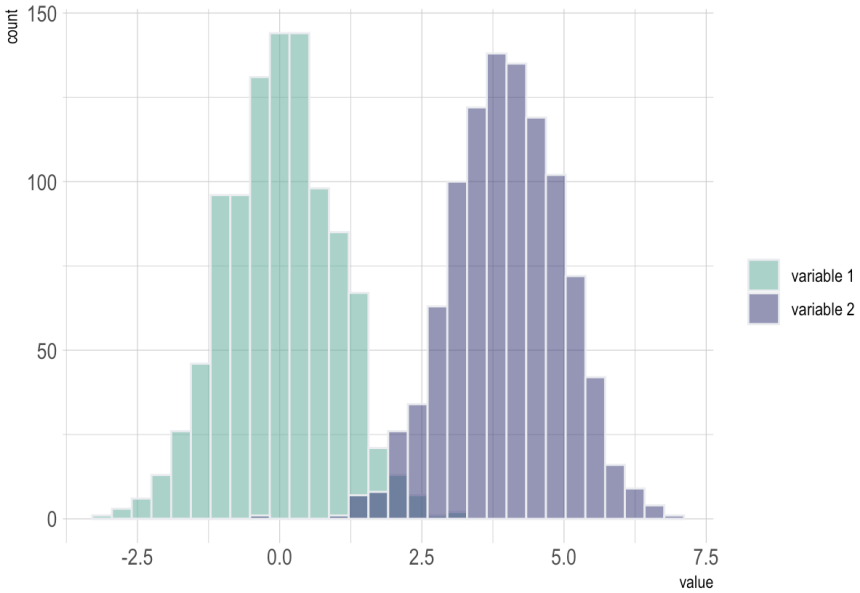
choose bin size

5

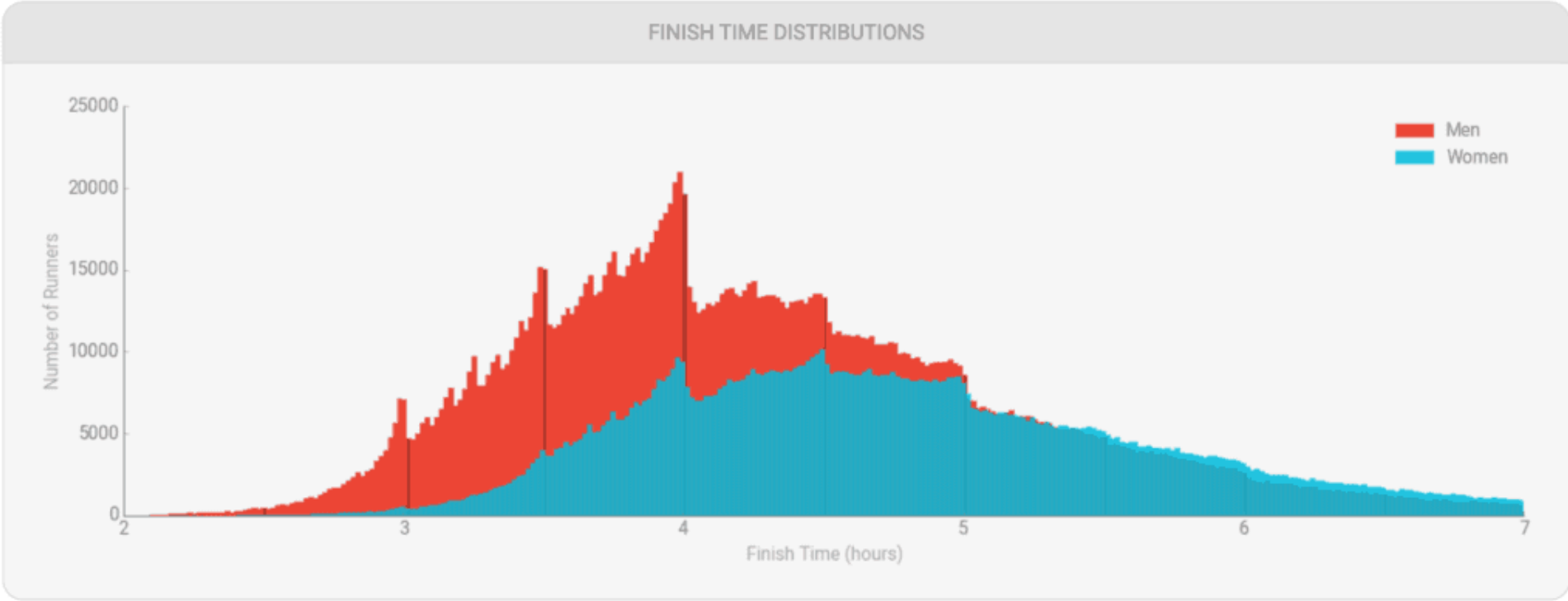

Number of occurrences
0
20
40
60
80
100
10
15
20
25
30
35
40
45
50
55
60
65
70
Value bins

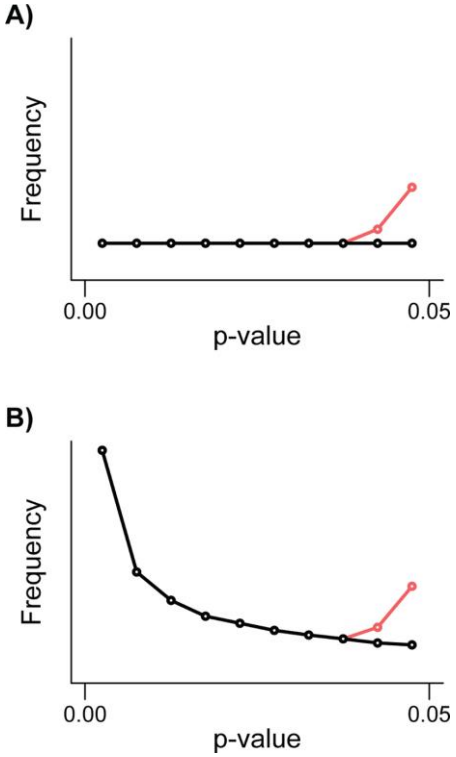

Symmetric (normal) vs skewed and uniform distriutions
Normal distribution
(unimodal, symmetric,
the "bell curve")
Right-skewed
distribution
(Positively-skewed)
Left-skewed
distribution
(Negatively-skewed)
Uniform distribution
(equal spread,
no peaks)
Unimodal vs bimodal distributions
Normal distribution
(unimodal, symmetric,
the "bell curve")
Symmetric bimodal
distribution
(two modes)
Non-symmetric
bimodal distribution
(two modes)


Raw Data
Box-plot of the Data
Violin-plot of the Data
A B C D E F G
A B C D E F G
A B C D E F G


count
150
100
50
0
-2.5
0.0
2.5
5.0
7.5
value
variable 1
variable 2

University of Exeter

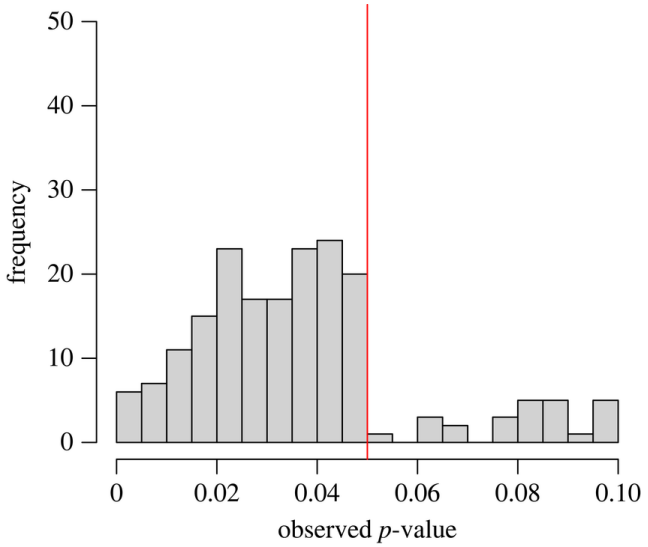




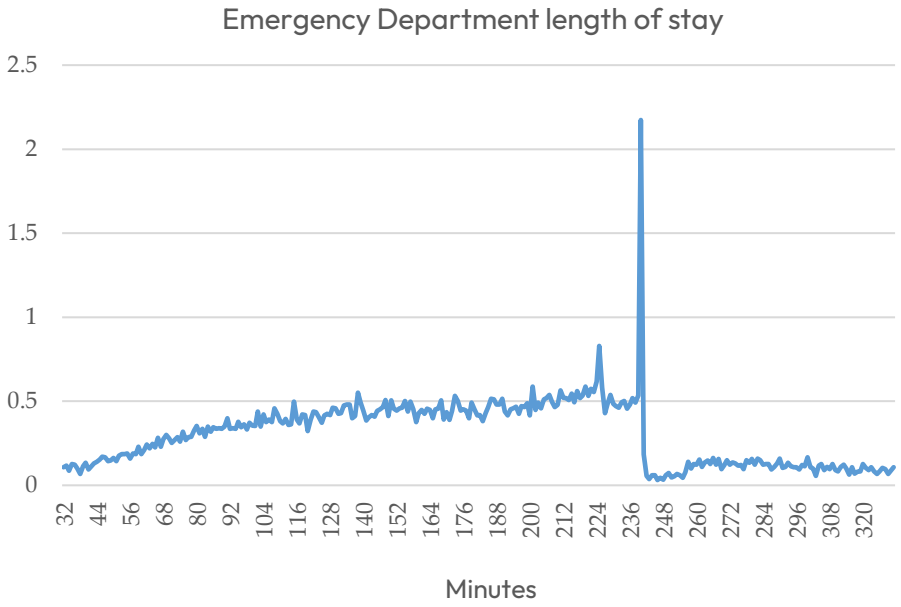
P-hacking is a bias in the scientific literature that occurs when researchers manipulate data or statistical analyses to obtain significant results.

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

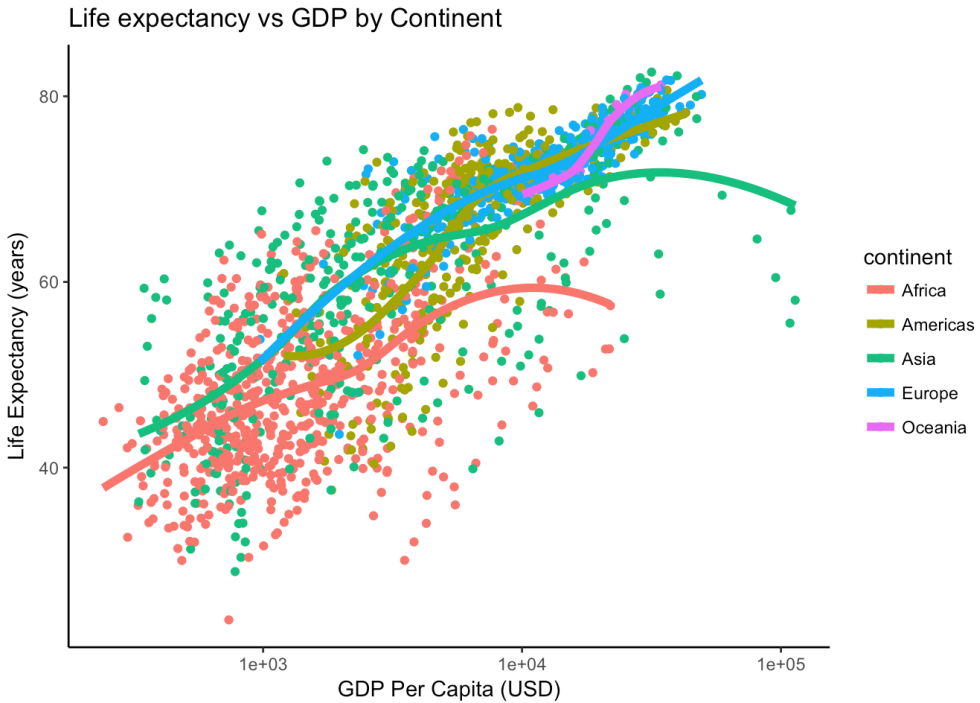
P-hacking can lead to false or exaggerated findings and undermine the validity of scientific research.



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.



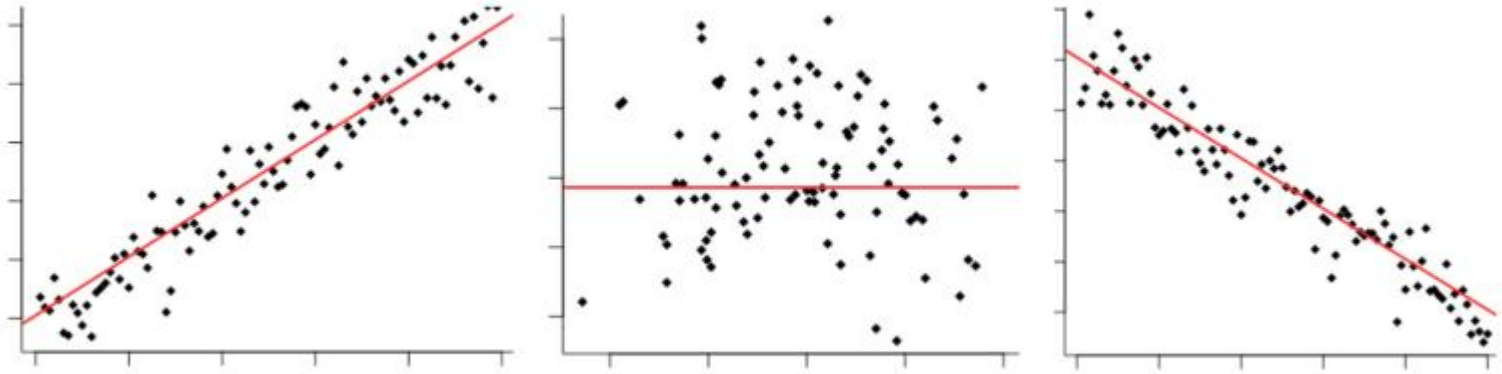
Can anyone guess what is happening in this histogram?



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

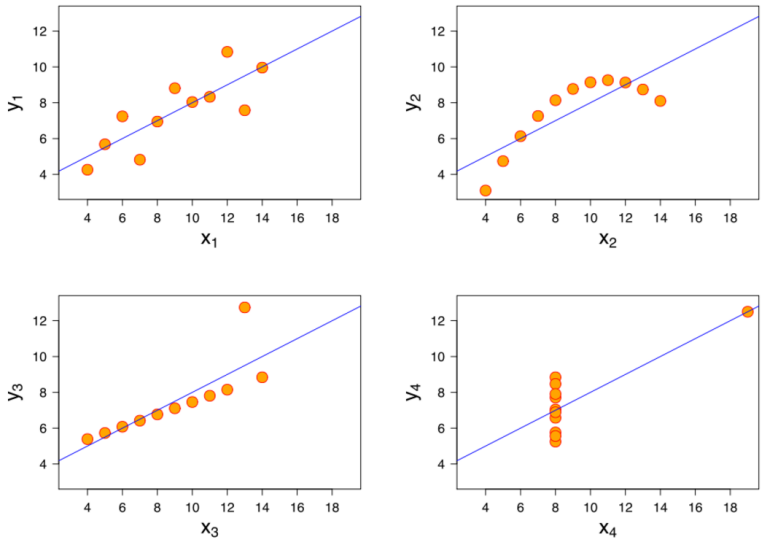
I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0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

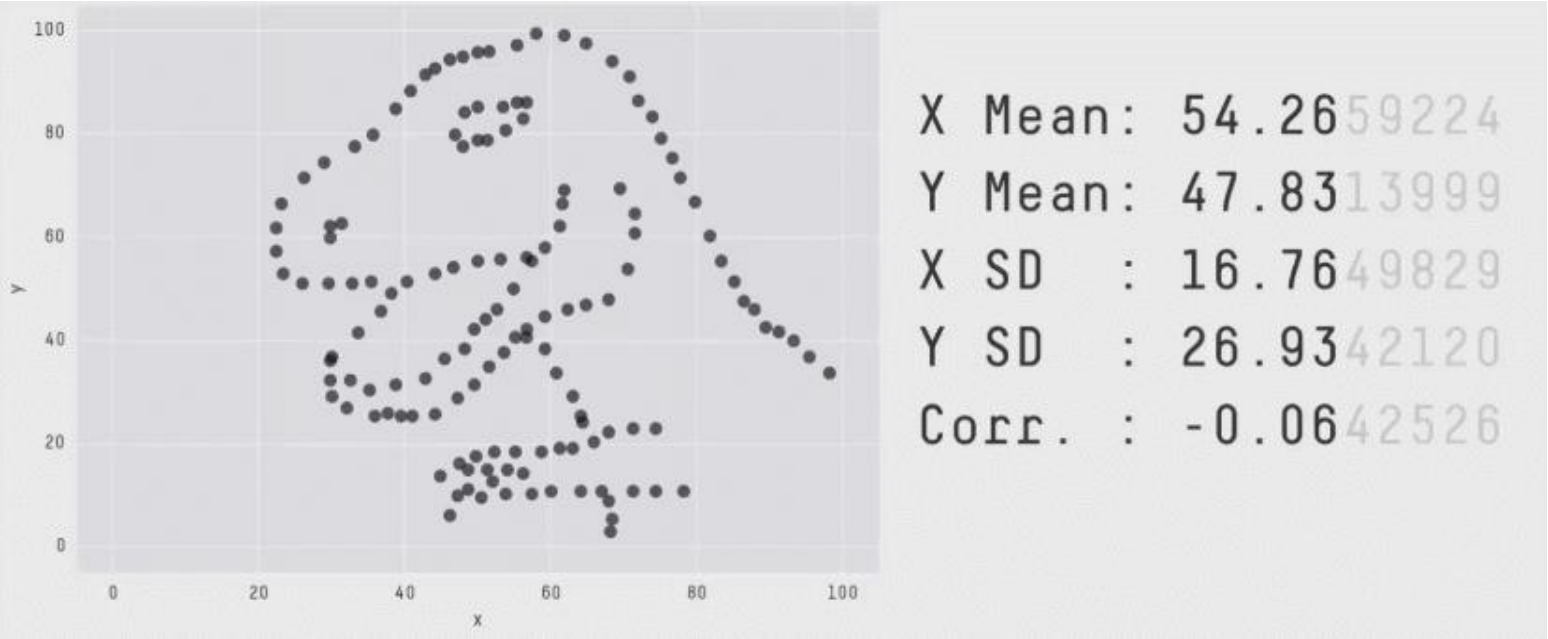
Source

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)

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





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.

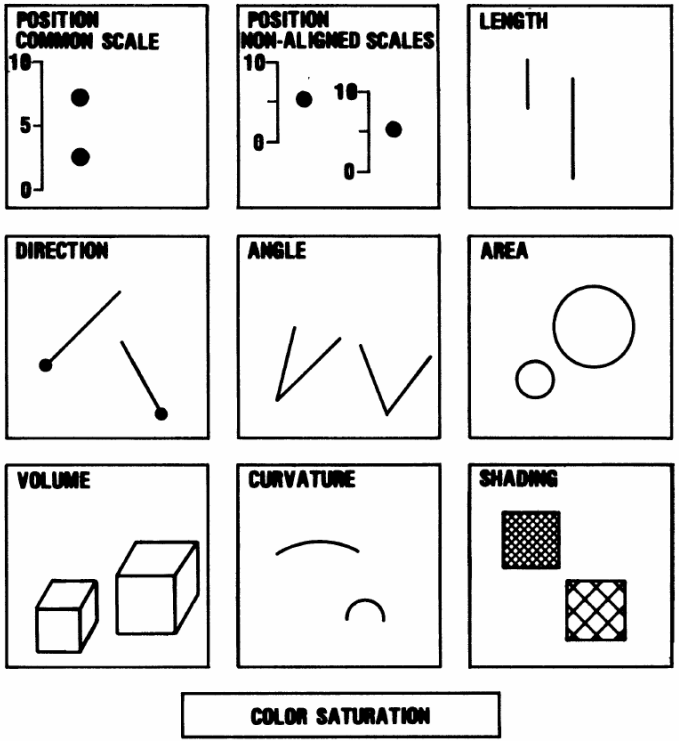
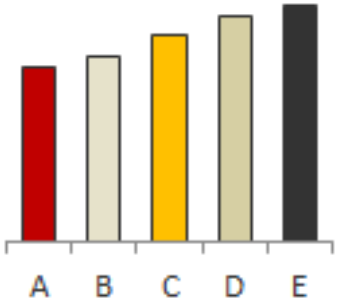
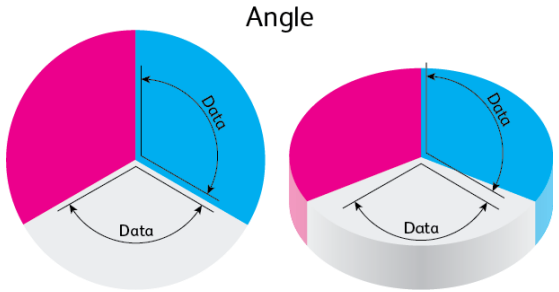


Figure 1. Elementary perceptual tasks.

Why do people hate Pie Charts?

Pie charts aren't *bad* visualizations. They just need to be used appropriately.

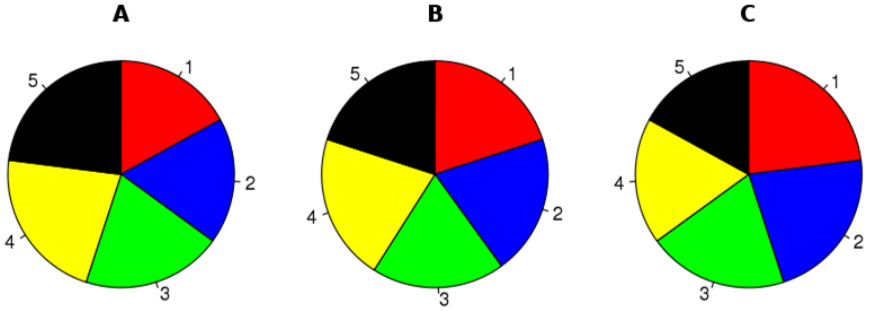




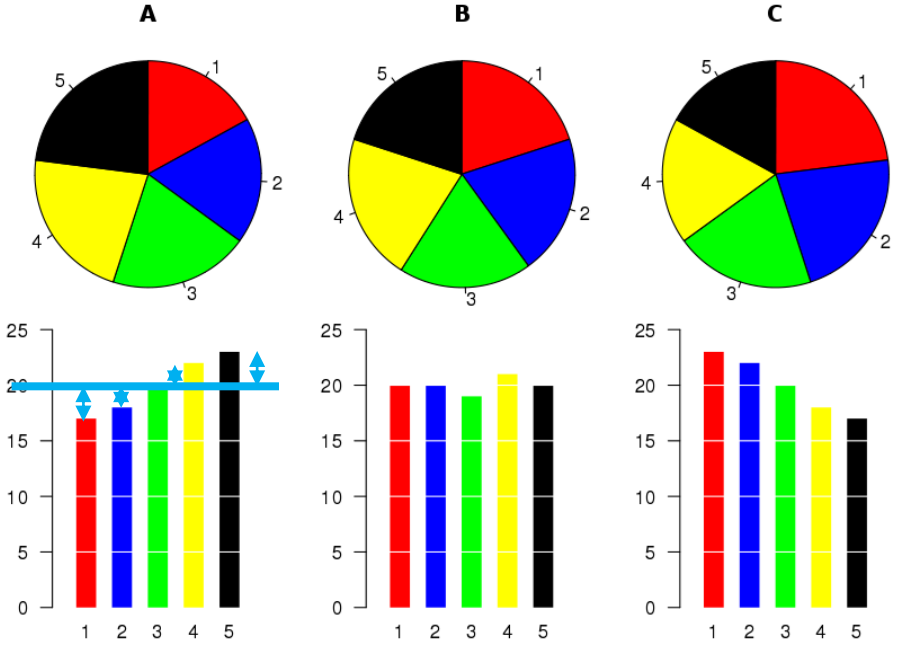
The data in a pie chart is encoded in **the angle** of the slices.

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….



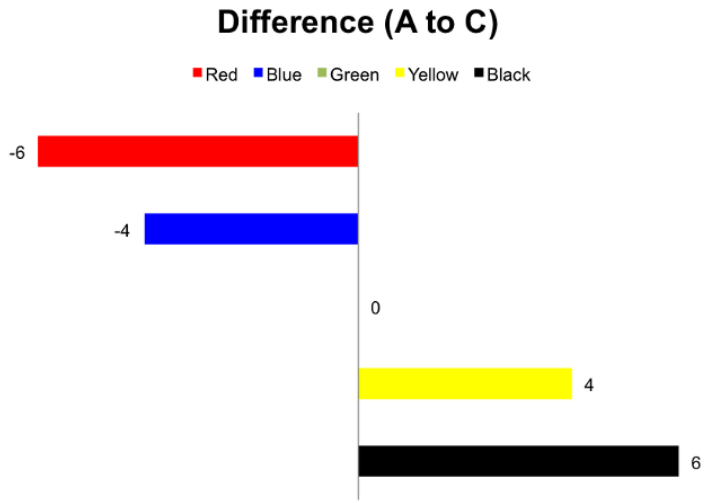
Looking at Pie Charts....



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?

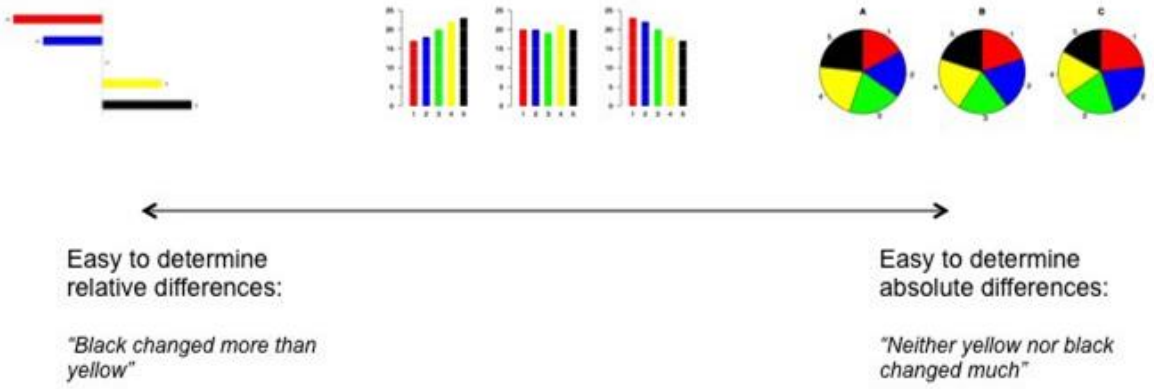
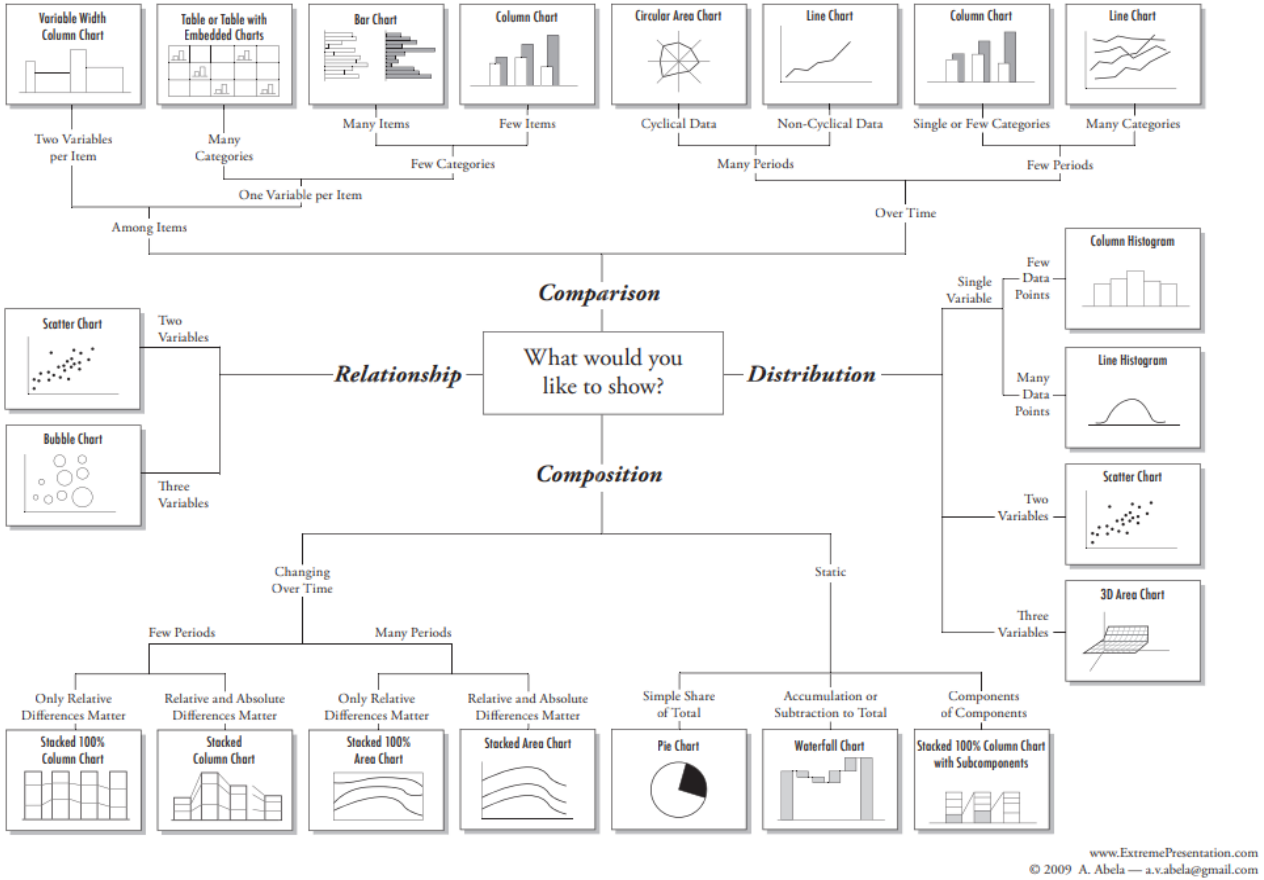
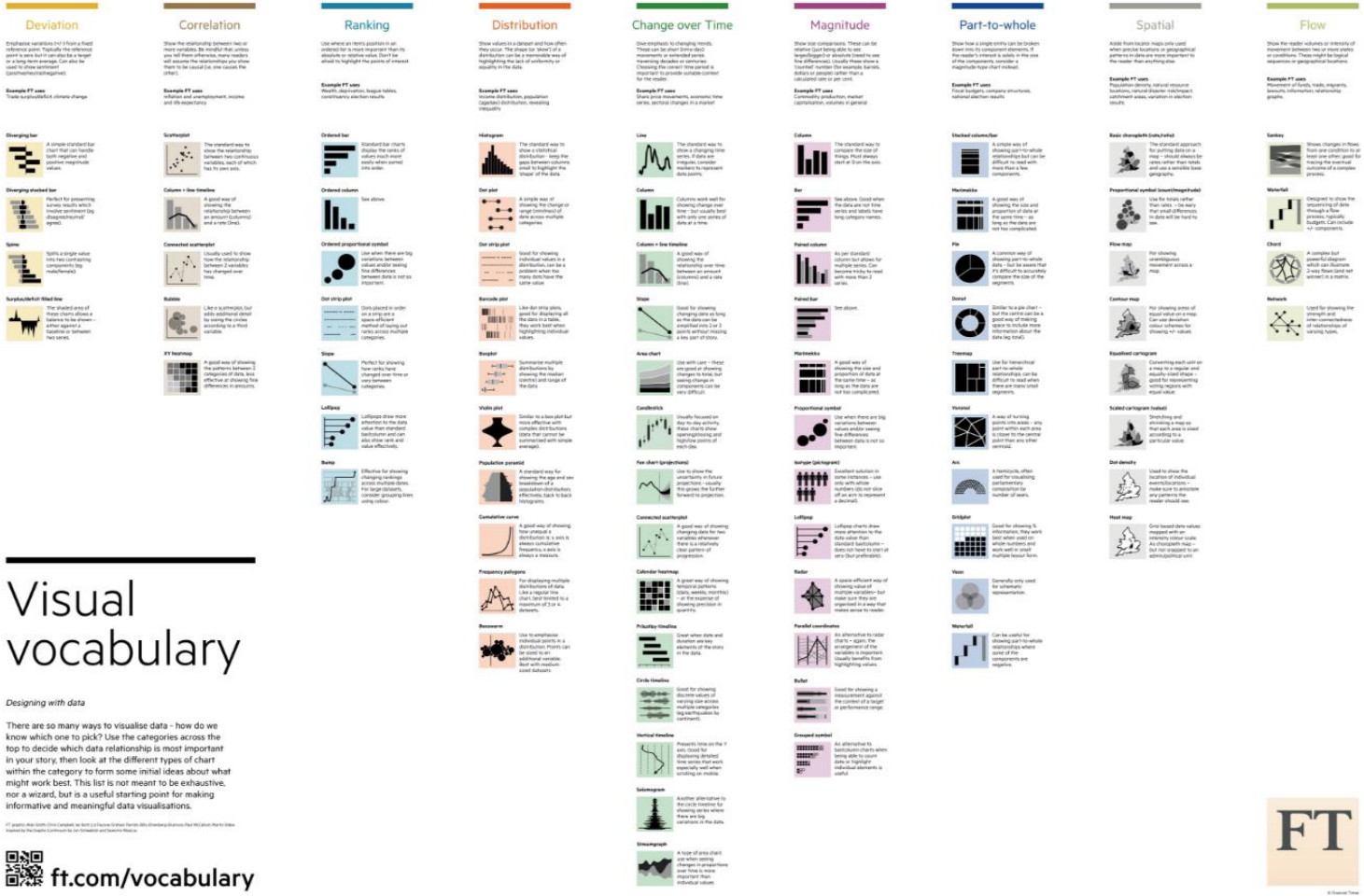


Chart Suggestions—A Thought-Starter



Reference: The Chart Chooser



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.

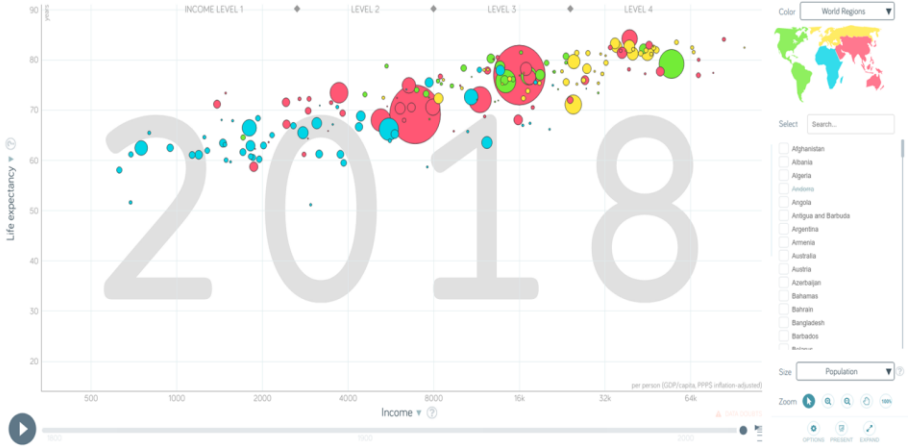
Source.

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.

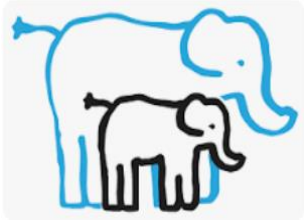
Gapminder has continued on without Hans' input since his passing and continued to help put world development and human experience data in context.



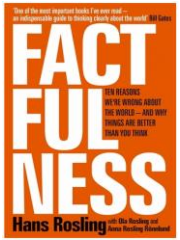
University of Exeter

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

Ink to Data Ratio

- The "data-ink ratio" : Edward Tufte, a prominent expert in data visualisation.

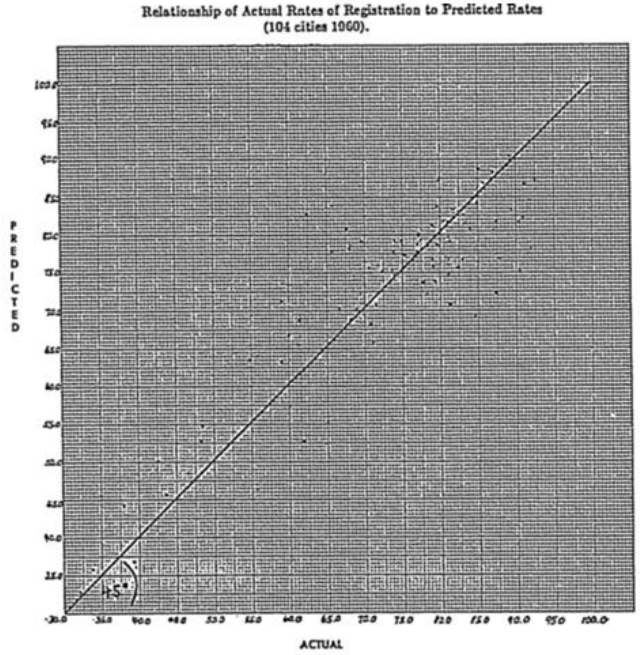
- It refers to the proportion of ink (or pixels in digital formats) used in a visualisation that is directly related to representing the data, as opposed to ink used for labels, decorations, or non-essential elements.

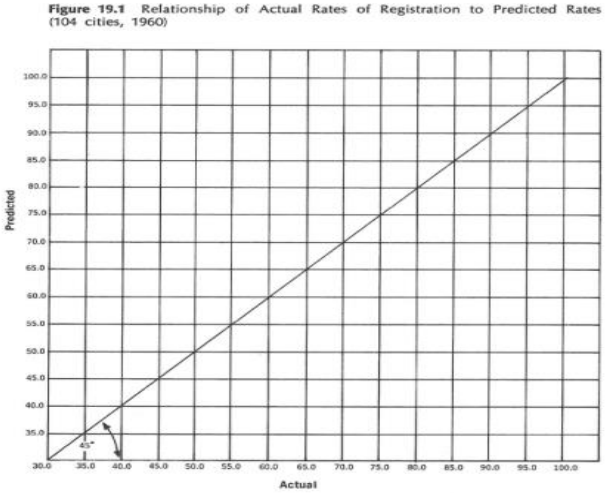
- It encourages the minimisation of unnecessary ink to maximise efficient and clear visualisations, remove clutter, redundant, or distracting elements.

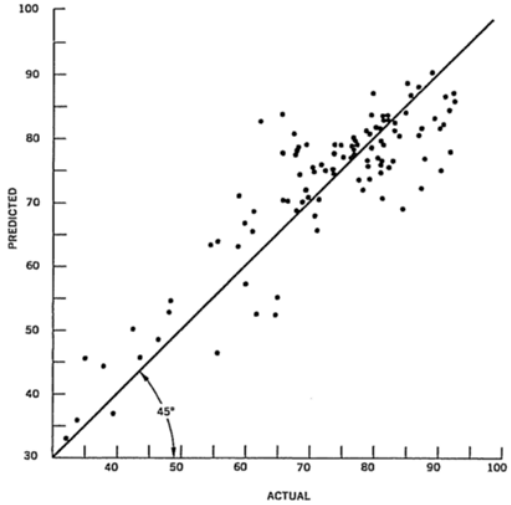
- It encourages designers to concisely convey the information the visualization is intended to communicate. This helps viewers quickly grasp the main message and insights from the data.

Principles of Data Ink

Above all else show data.

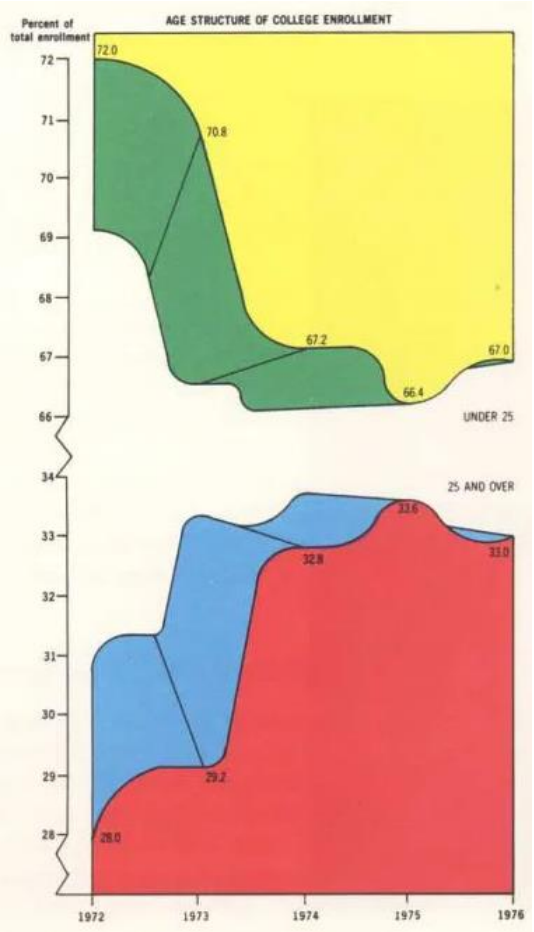






Relationship of Actual Rates of Registration to Predicted Rates (104 cities 1960).

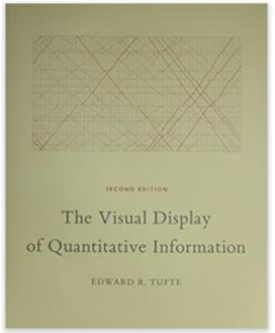
Source: (Tufte, 2001)



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, 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."*

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

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.

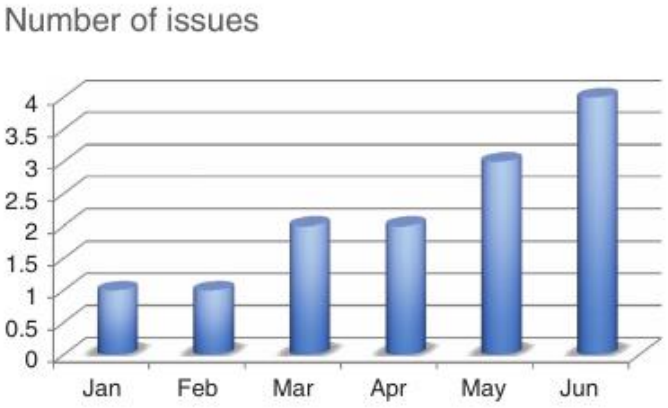
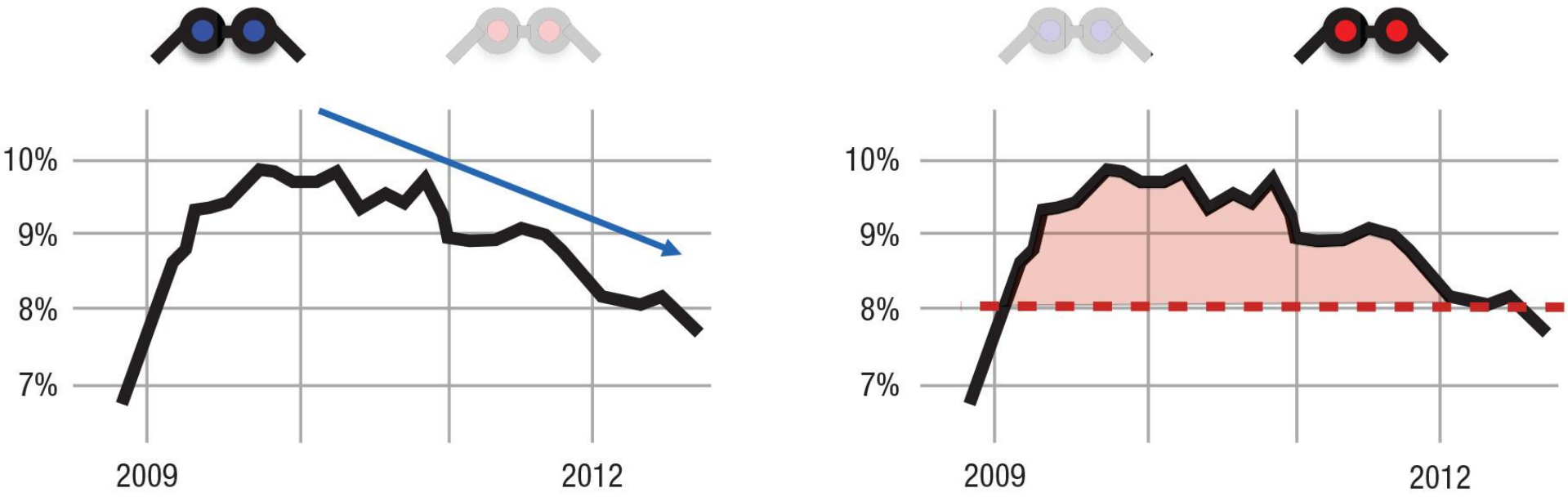


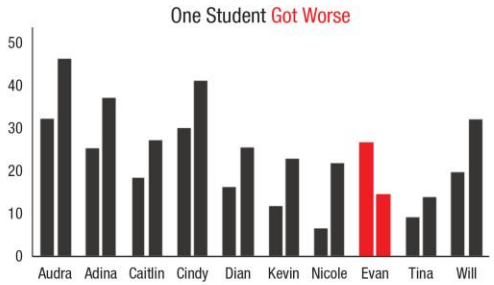
FIGURE 2.25 3D column chart

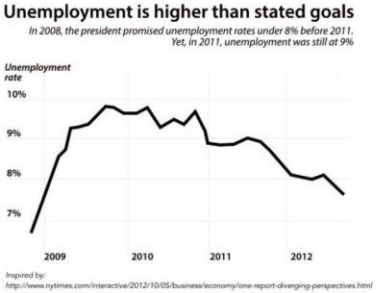

Remove
to improve
(the data-ink ratio)
Created by Darkhorse Analytics
www.darkhorseanalytics.com

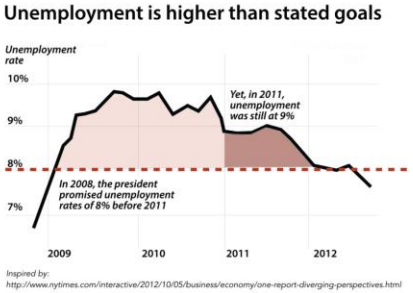


An example of emphasizing different perspectives in a single data set (inspired by Bostock et al., 2012). One data set can be seen with dramatically different perspectives, depending on which patterns an observer does and does not extract.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), 110-161.







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.

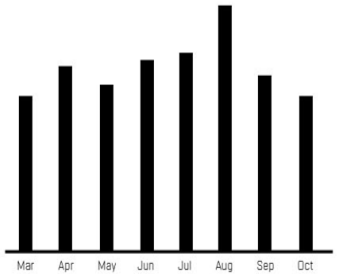
The graphic at the top illustrates a colour-highlighting technique suggested in business-oriented practitioner guides (e.g., Knaflic, 2015).

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

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), 110-161.

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.



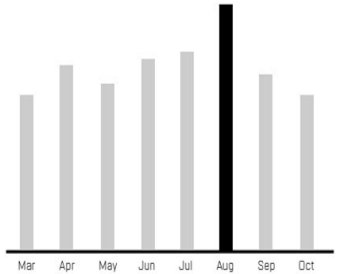
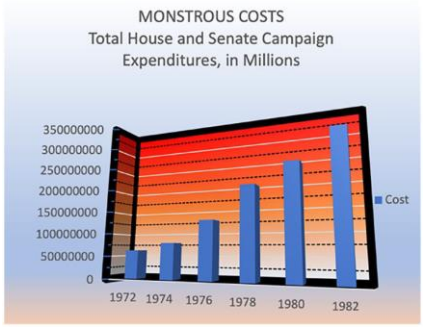
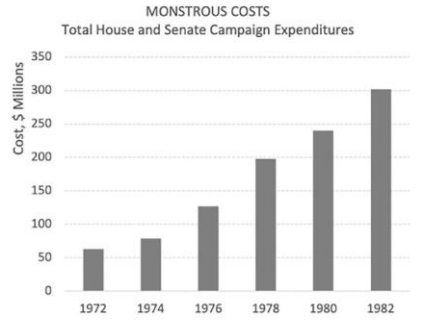
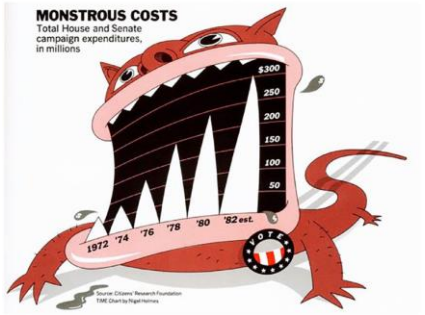


Image Source







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*.

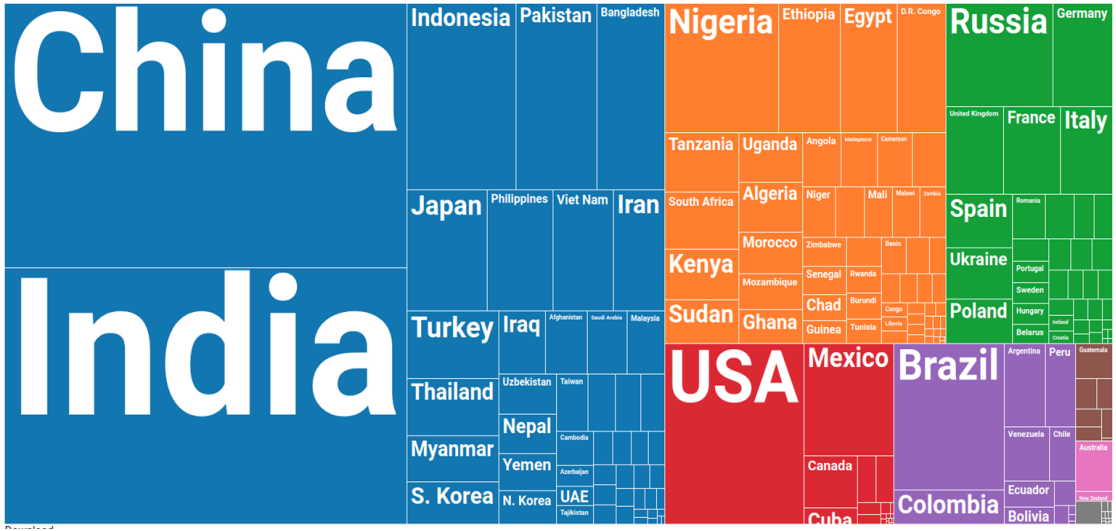
Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), 110-161.

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.

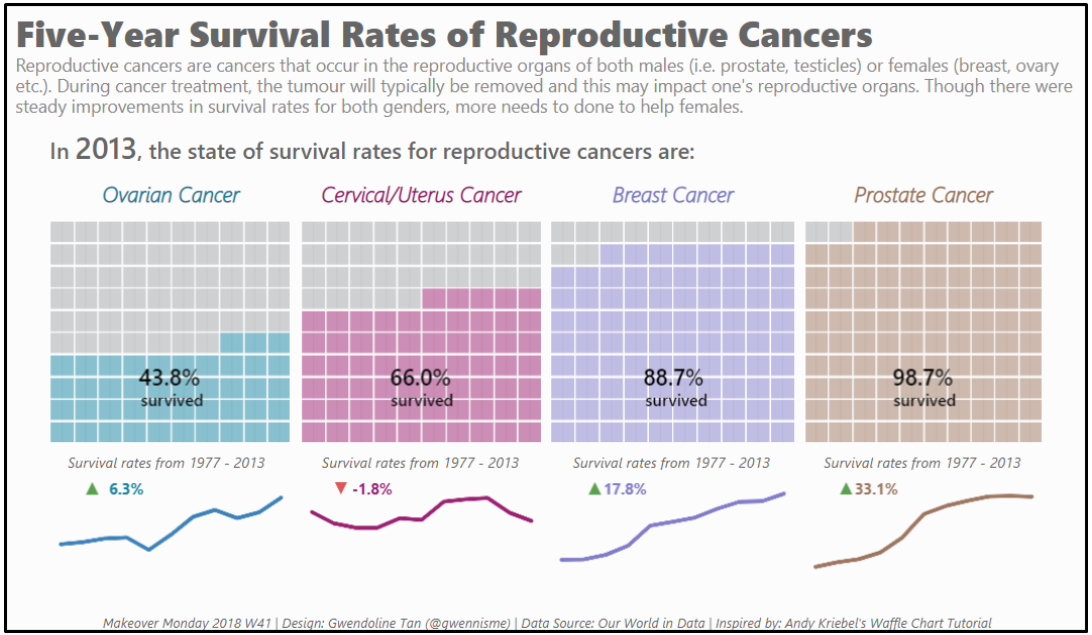
See this interactive map of population.



Proportions. Waffle plots.

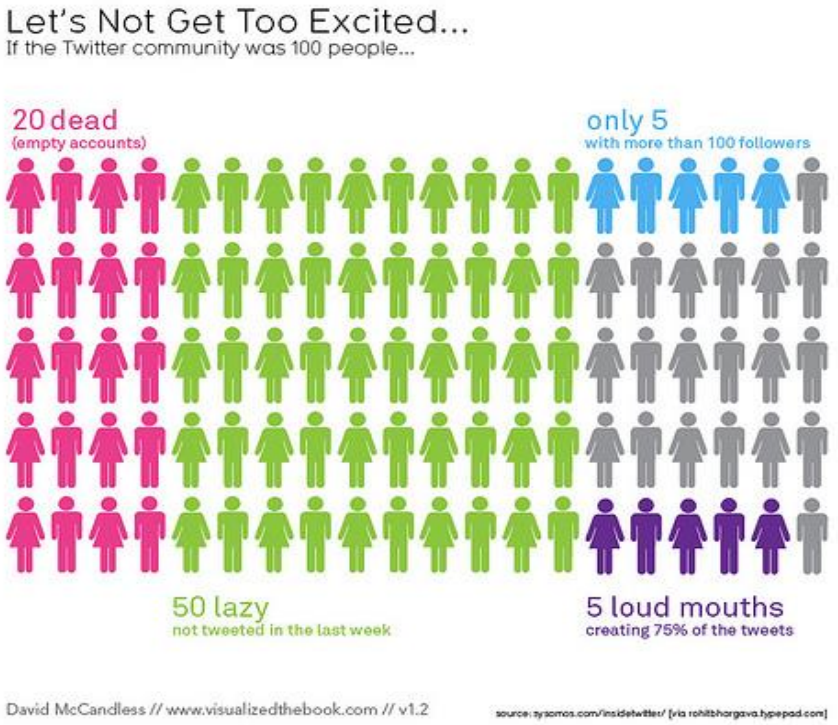
Five alternatives to the pie chart.

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



Proportions. Isotype/pictograms.

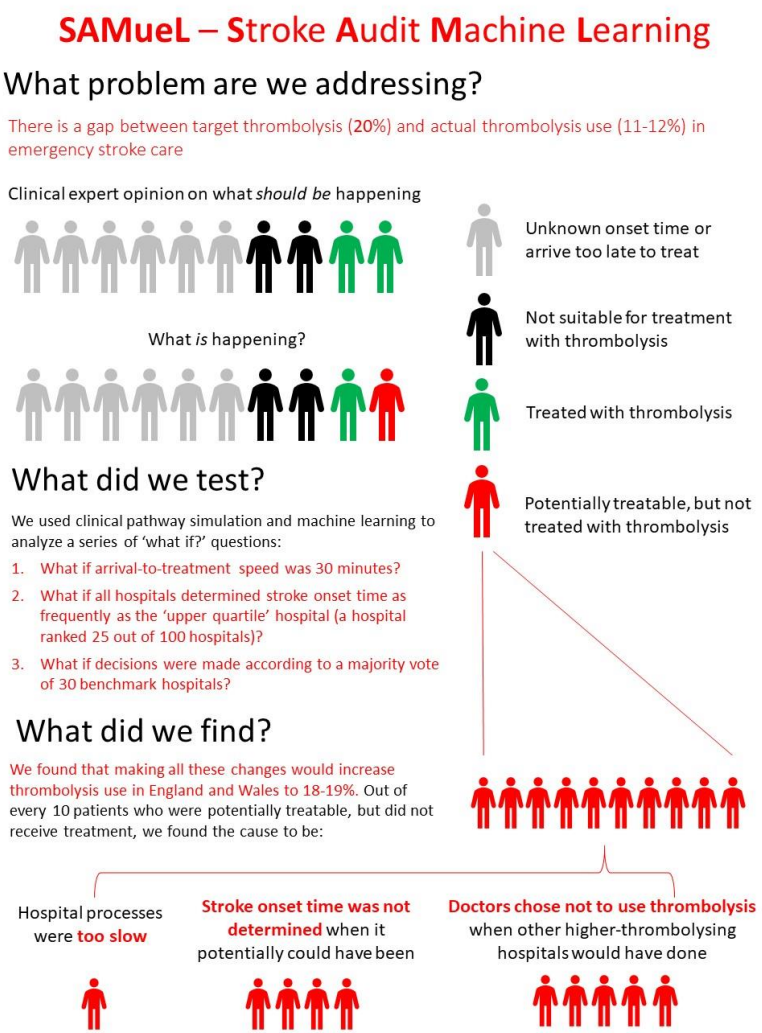
Reduce the population to 100 people. Then colour them proportionally.



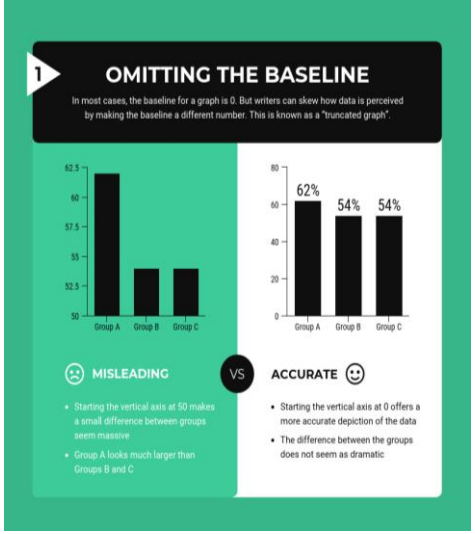
Proportions. Isotype/pictograms.

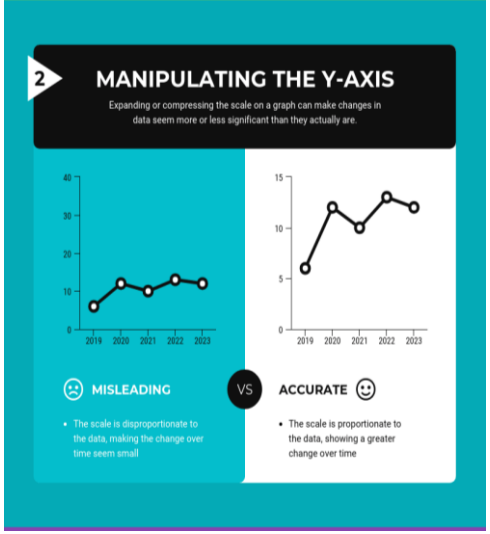
Reduce the population to 100 people. Then colour them proportionally.

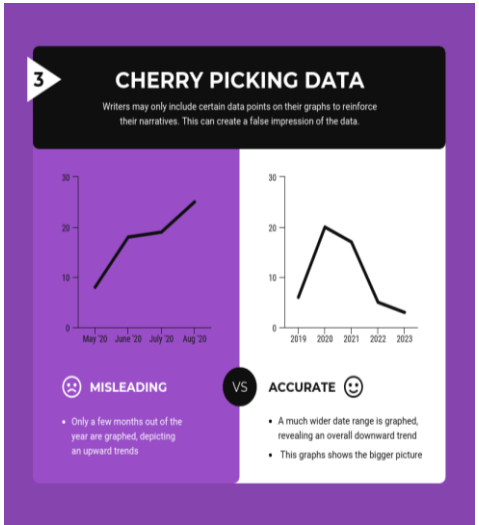
[Introduction to SAMueL-1 (Stroke Audit Machine Learning) — SAMueL Stroke Audit Machine Learning 1 (samuel-book.github.io)]

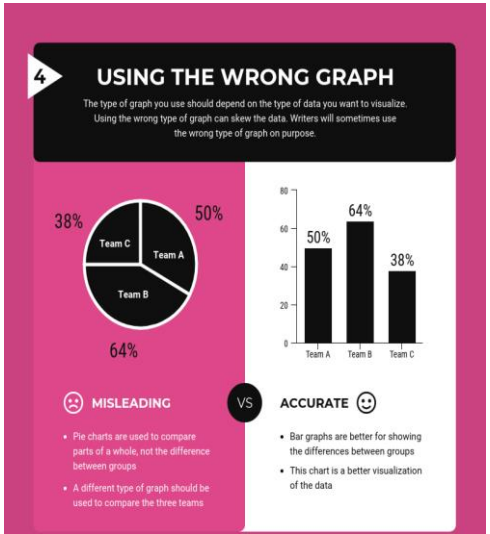


Using graphs to mislead

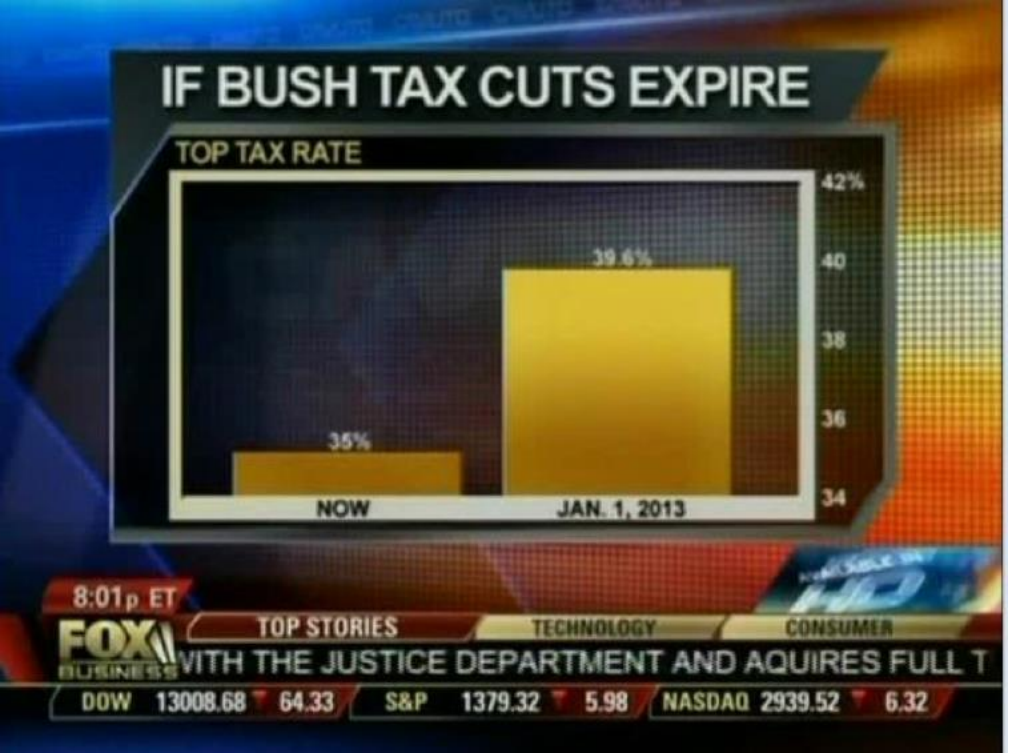


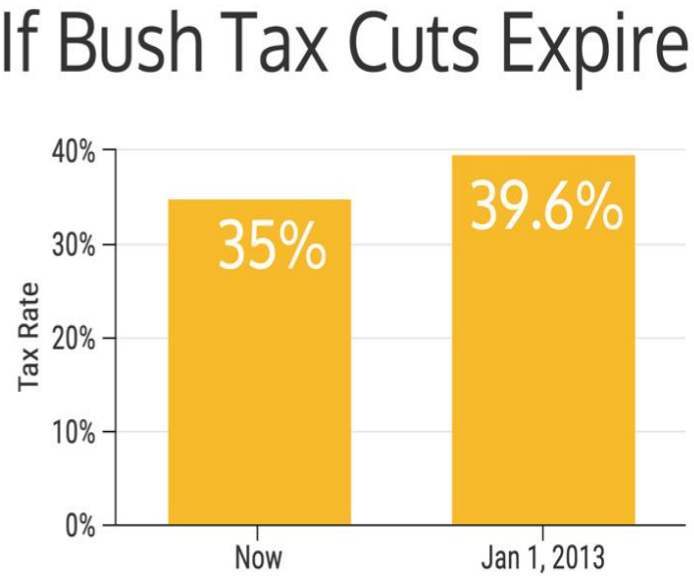






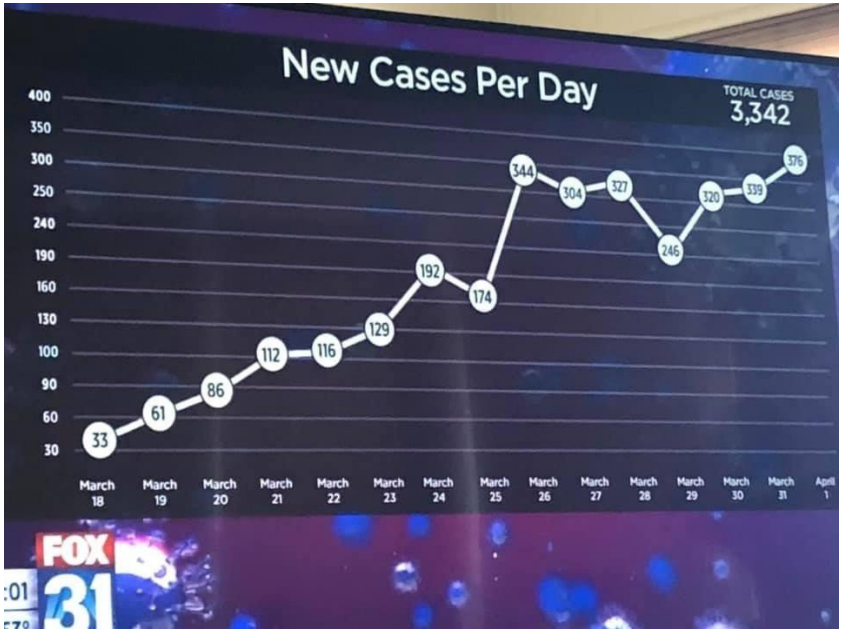
Source





Deceptive Designs

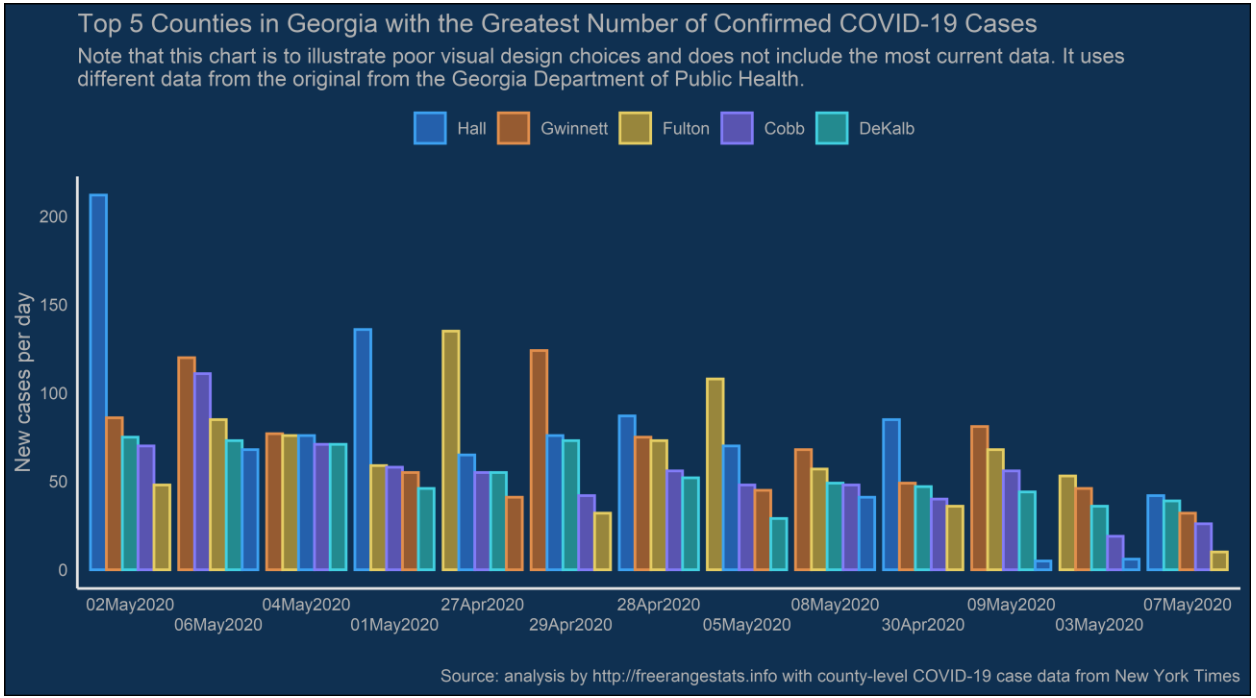
Deceptive Design



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



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

Highest

Position

1 7

Length

1 7

Area

1 7

Angle

1 7

Intensity

1 7

Lowest

Common Illusions That Distort Data

Caveats for the visual encoding in each row

99 98 a b 99 98 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.

1 7

But beware: if the number is actually mapped to *Length*, the gray circles depict ~1:2.5 ratio.

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

a b c d e f

Max height
Mean height
Min height

Stacked Bar

a b c d e f

Min
Max
Mean length of dark bars

Bubble Map

Mean Area
Min
Max

Slope Graph

Max
Mean Angle
Min

Heat Map

Min Max

Mean Intensity

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

Δ#

a b c d e f

#

a b c d e f

"a, c, & e have increased"

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

a b c d e f

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.

a b c d e f

For color heat maps, depict deltas as blue (+) & red (–)

[green/red is unsafe for colorblindness]

a b c d e f

University of Exeter

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.

Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the public interest*, *22*(3), 110-161.

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?

?