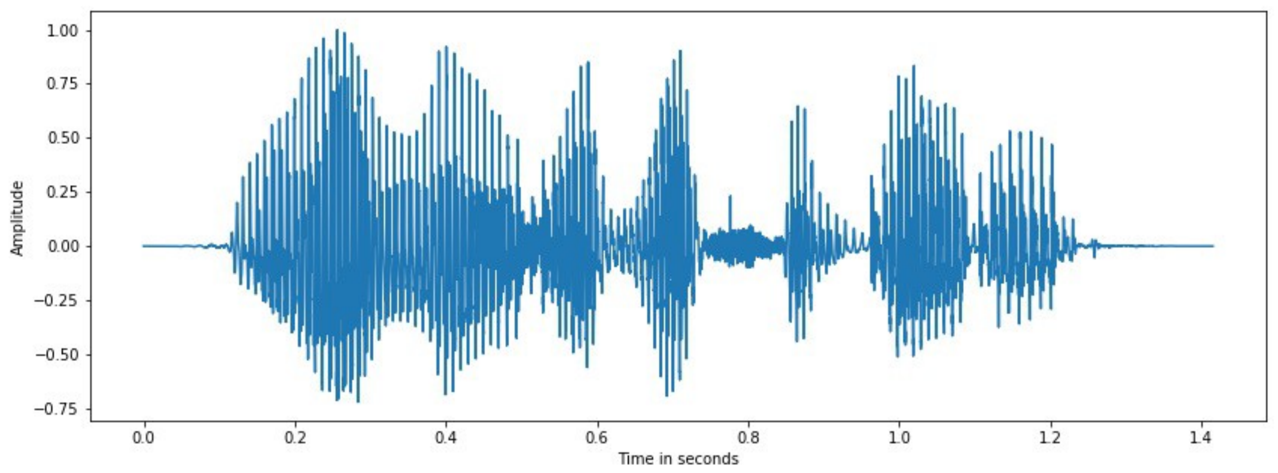
Data Analysis and Machine Learning 4

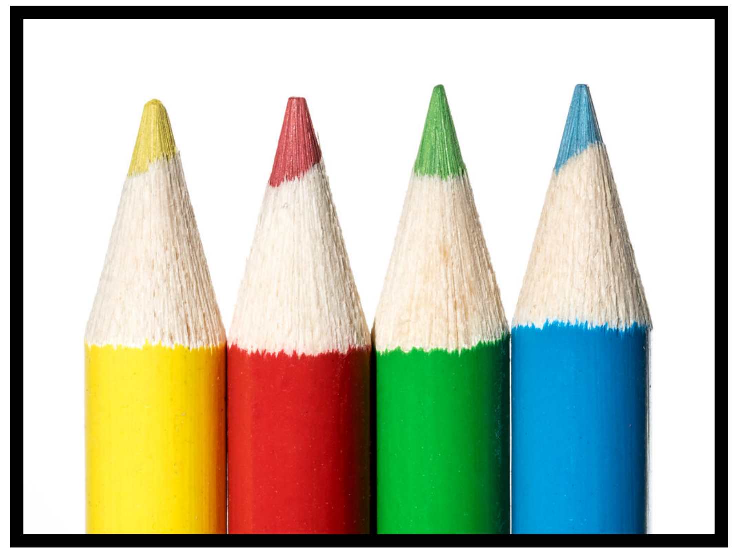
Week 2: Summarising and visualising data

Elliot J. Crowley, 23rd January 2023

Recap

- We looked at different modalities of data





	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

- We considered variable types

iris species (nominal)

level of education (ordinal)

Tabular data

- We will focus on this modality quite a bit
- It crops up a lot in real life and it is straightforward to work with

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	virginica
146	6.3	2.5	5.0	1.9	virginica
147	6.5	3.0	5.2	2.0	virginica
148	6.2	3.4	5.4	2.3	virginica
149	5.9	3.0	5.1	1.8	virginica

Summarising Data

World Happiness Report

- Produced by a non-profit of the United Nations
- What do you want to know when you see this?

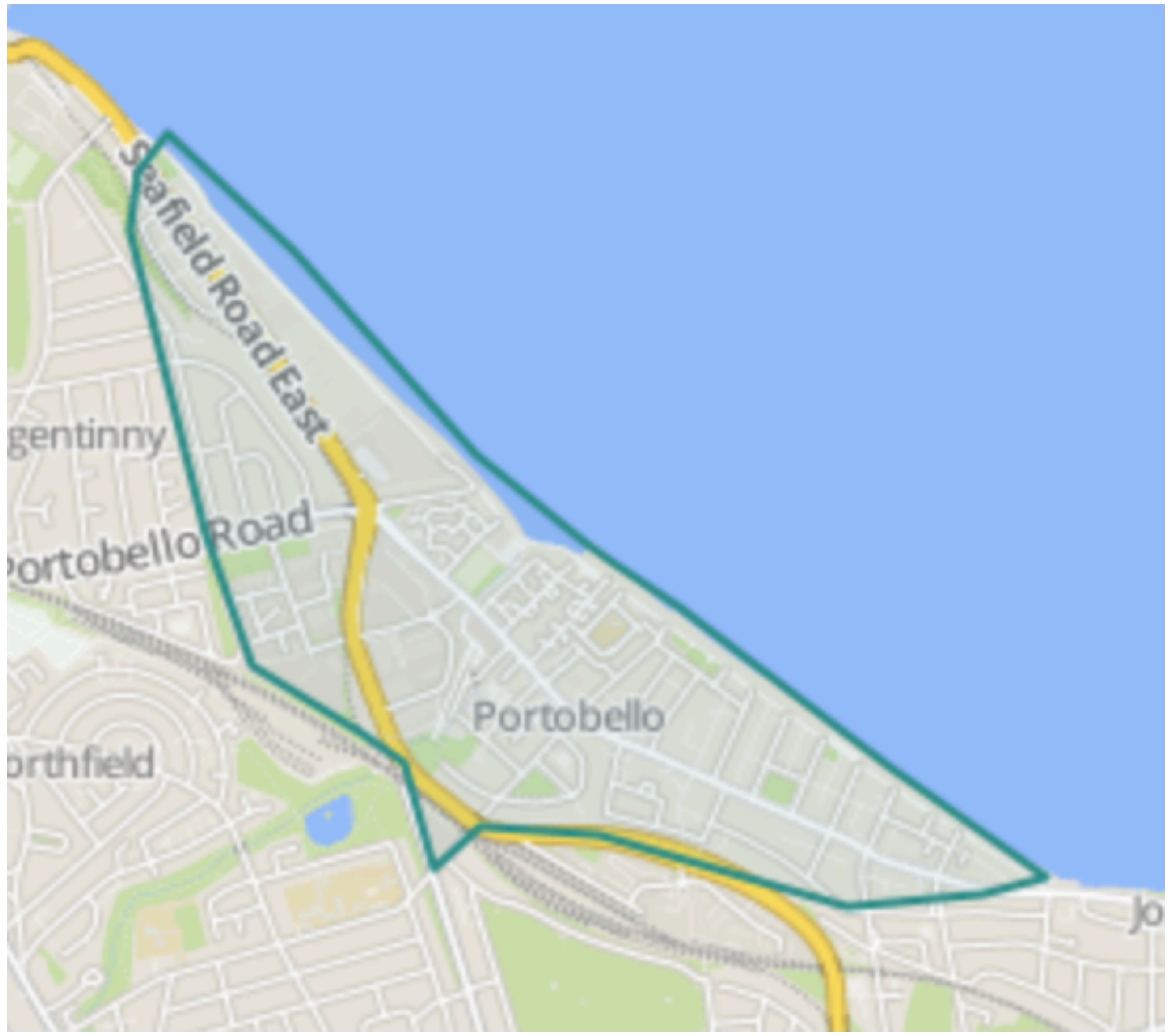
Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Guatemala	6.436	0.800	1.269	0.746	0.535	0.175	0.078
Yemen	3.380	0.287	1.163	0.463	0.143	0.108	0.077
Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
...	...	...	...	...	...	...	...
Libya	5.525	1.044	1.303	0.673	0.416	0.133	0.152
Jamaica	5.890	0.831	1.478	0.831	0.490	0.107	0.028
United States	6.892	1.433	1.457	0.874	0.454	0.280	0.128

Extreme values

- Take **maximum** of score: Finland
- Take **minimum** of perceived corruption: Moldova

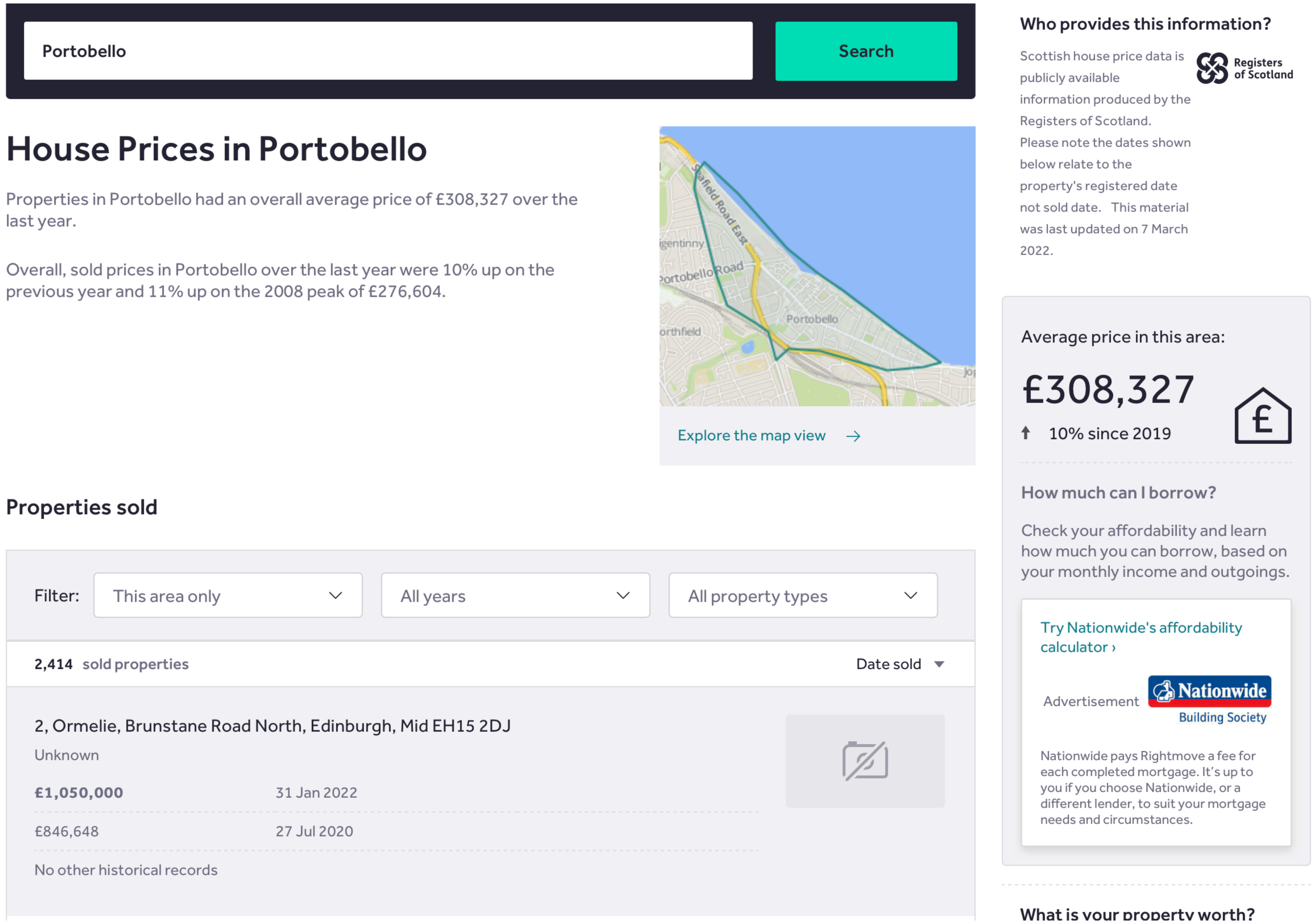
House buying

- Let’s say I’m considering buying a property in Portobello
- What do I need to know?



Central values

- Good to know the mean house price
- Or median?



Source: Rightmove

Summary Statistics

- Most people will not scroll through a table!
- Summary statistics let us convey information as simply as possible

WORLD ›

99% of the world is breathing poor-quality air, WHO says

APRIL 4, 2022 / 3:01 PM / AP

Salaries in London Area

Location: – London Area | or | Find a Specific Employer: Employer's Name | Search | Sort: Popular

Company	Average Base Salary in (GBP)	Range
Accenture London 4.1 ★ 21 salaries See 21 salaries from all locations	£54,608 /yr	£32K – £102K
Deloitte London 4.0 ★ 19 salaries See 20 salaries from all locations	£58,219 /yr	£31K – £105K
Barclays London 4.0 ★ 16 salaries See 16 salaries from all locations	£52,872 /yr	£18K – £109K
University College London London 4.3 ★ 10 salaries See 10 salaries from all locations	£39,800 /yr	£18K – £57K

Source: Glassdoor

Mode

- Suitable for summarising ordinal, nominal, and discrete variables
- Let's denote our variable (e.g. iris species) X
- We have measurements of that variable
- The mode is the measurement that occurs the most

	Favourite Colour
0	red
1	blue
2	red
3	red
4	blue
5	yellow

3 red, 2 blue, 1 yellow

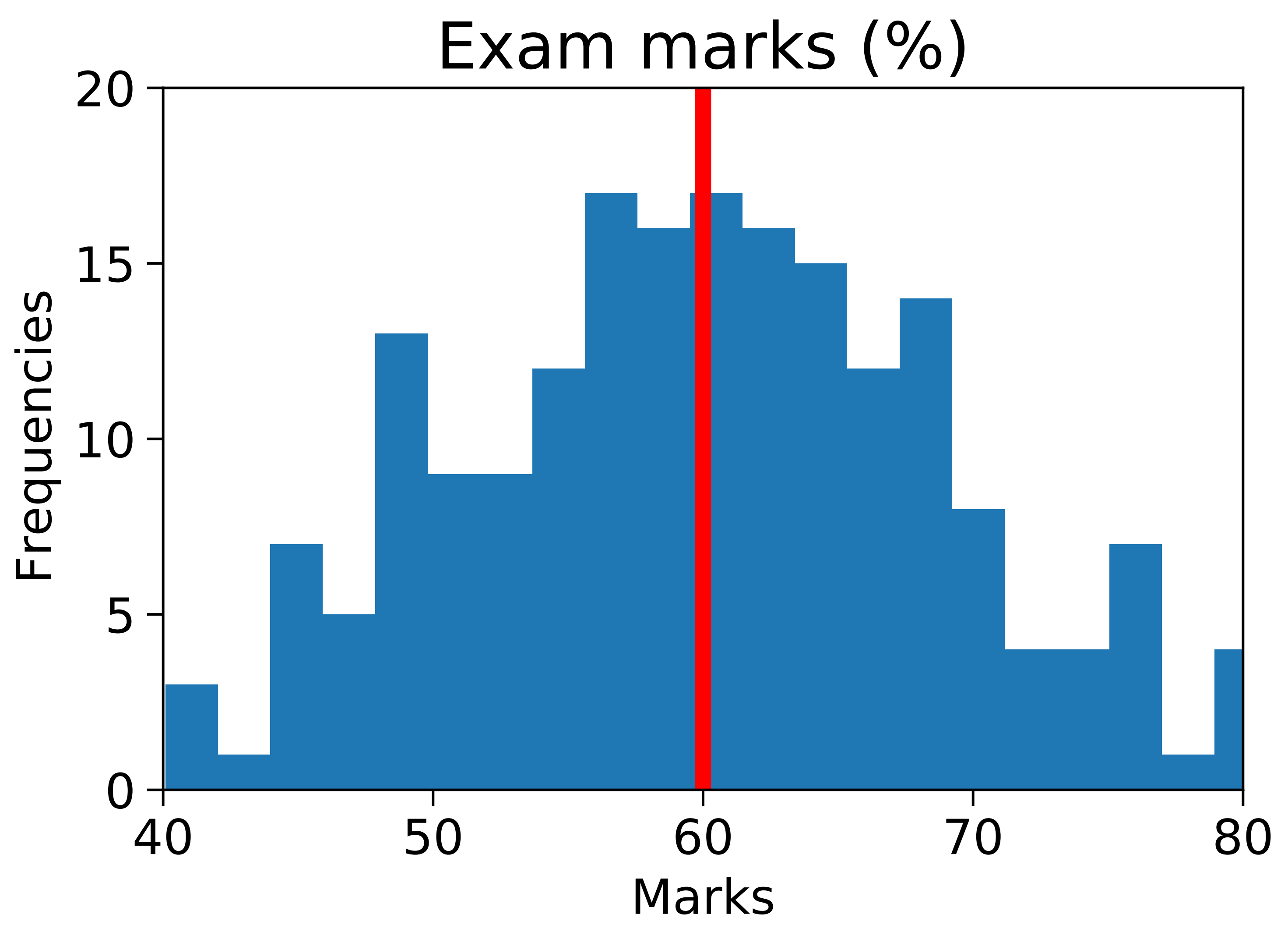
The mode is red

Mean

- Denote as μ. Suitable for summarising numerical variables
- For variable X we have N measurements $\{x^{(n)}\}_{n=0}^{N-1}$
- Counting from 0 because Python! Measurements are just $x^{(0)}, x^{(1)}, \ldots, x^{(N-1)}$

$$\mu_x = \frac{1}{N}\sum_{n=0}^{N-1} x^{(n)}$$

	Mark (%)
0	60
1	40
2	45
...	...



Variance and Standard Deviation

- Denote variance as σ^2. Standard deviation (SD) is σ
- For variable X we have N measurements $\{x^{(n)}\}_{n=0}^{N-1}$

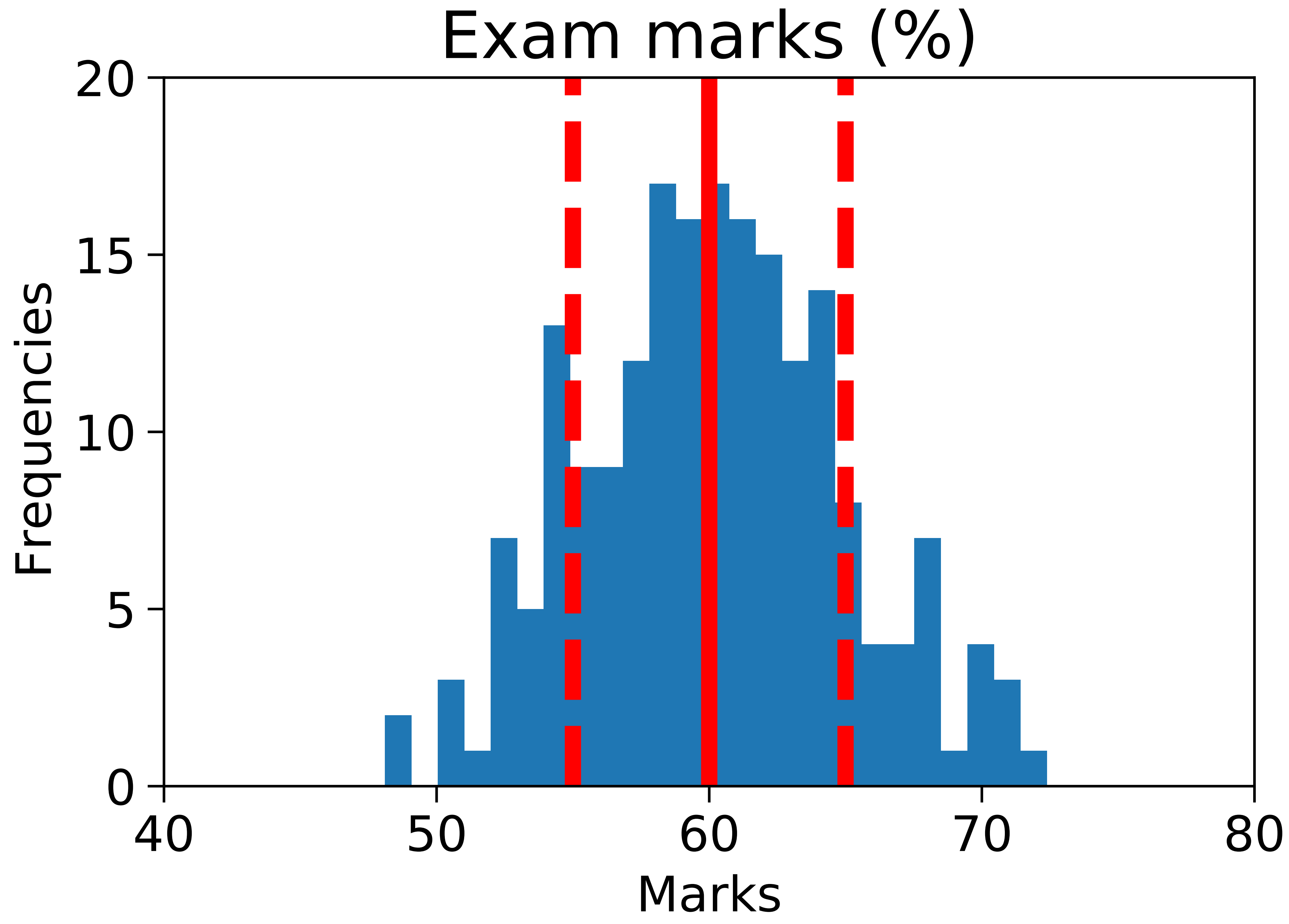
$$\sigma_x^2 = \frac{1}{N}\sum_{n=0}^{N-1}(x^{(n)} - \mu_x)^2$$

- Be aware that some definitions divide by $N - 1$
- $N \approx N + 1$ for large N so this isn't that important!

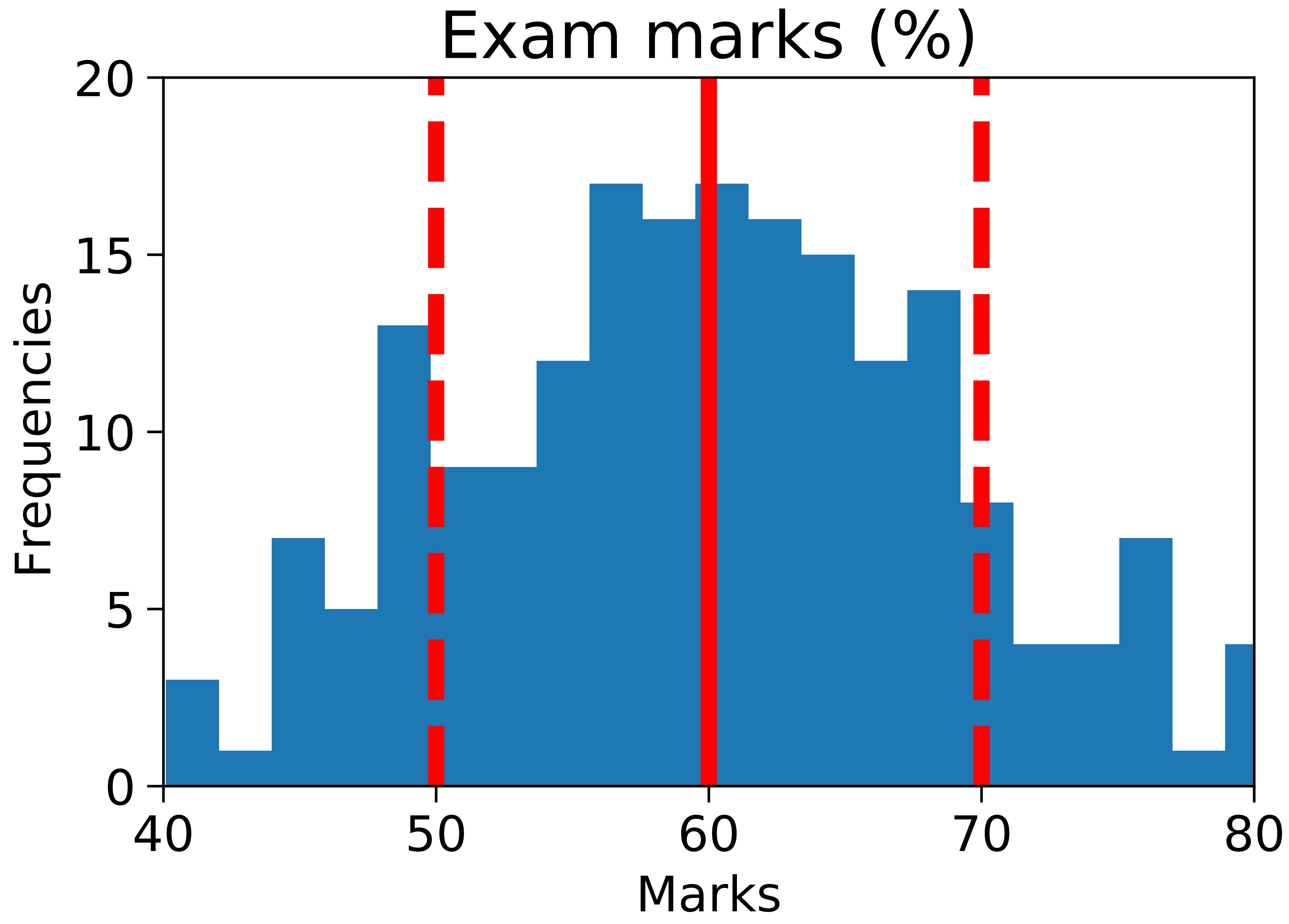
See https://towardsdatascience.com/the-reasoning-behind-bessels-correction-n-1-eeea25ec9bc9 for more info

Standard Deviation

SD measure the extent to which measurements deviate from the mean



$\sigma = 5$

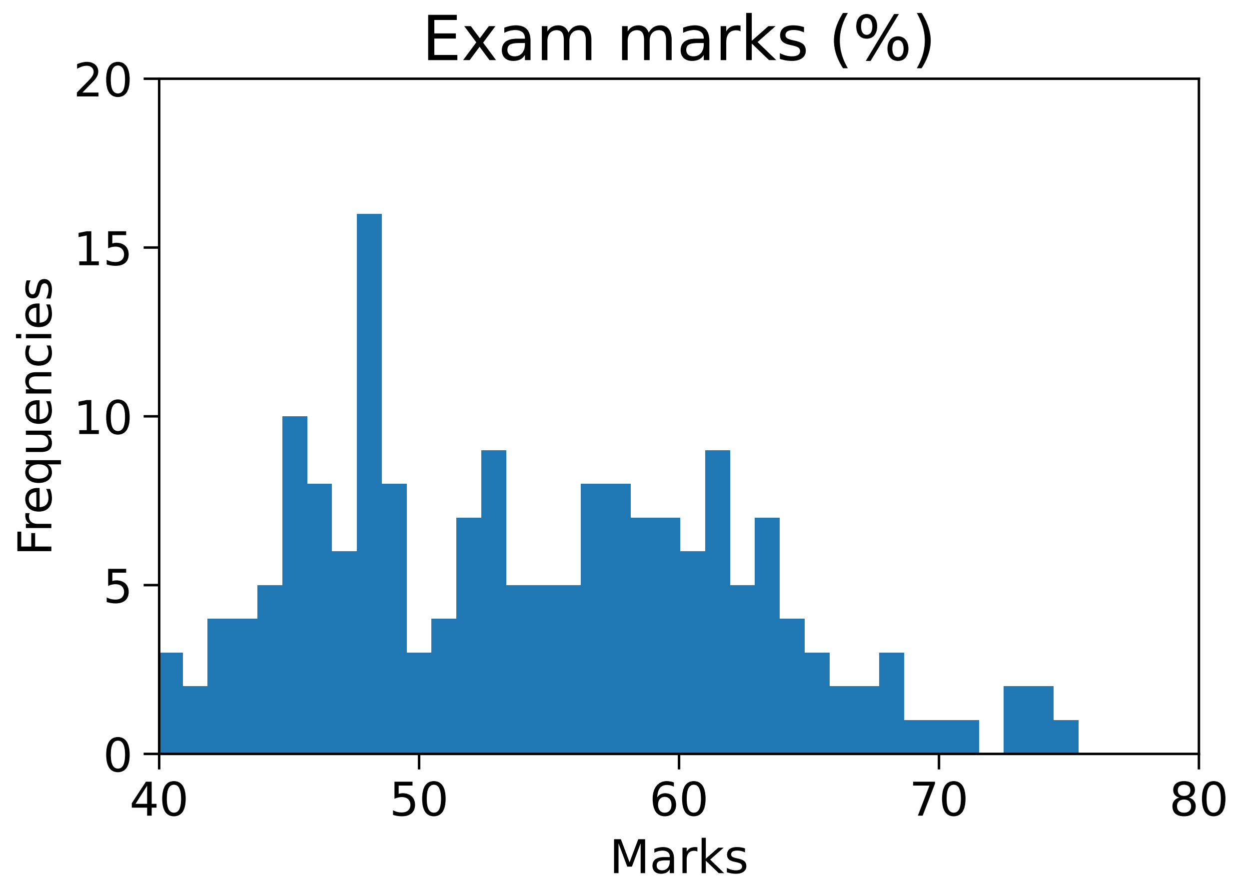


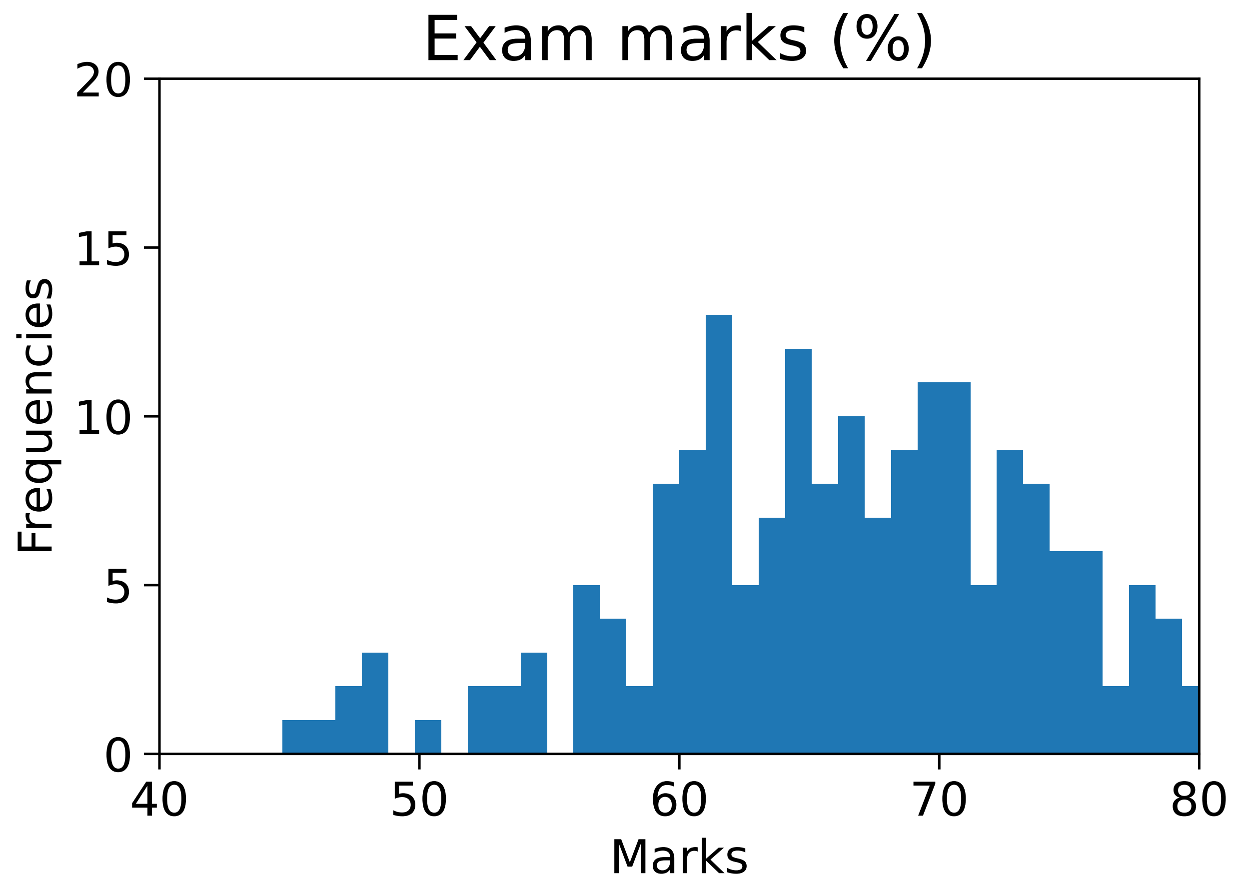
$\sigma = 10$

Skewness

- Denote using s. For variable X we have N measurements $\{x^{(n)}\}_{n=0}^{N-1}$

$$s_x = \frac{\frac{1}{N}\sum_{n=0}^{N-1}(x^{(n)} - \mu_x)^3}{\sigma_x^3}$$





Positive skew
Bulk of measurements on the left
Tail on the right

Negative skew
Bulk of measurements on the right
Tail on the left

Median

- Order measurements of a numerical variable from lowest to highest
- The median is the measurement in the middle

1 2 3 **5** 8 12 17

- The median is a **robust statistic**

1 2 3 **5** 8 12 1700000000

Medians are robust to outliers

Median salary is more meaningful than mean salary

Bet365 boss Denise Coates gets £300m pay package - a £170m cut

By Russell Hotten
BBC News

3 March

Denise Coates was appointed CBE for services to the community and business in 2012

Bet365 boss Denise Coates took home about £300m during its last financial year - £170m down on the previous year - as growth stalled.

Comments

BUSINESS

CEO pay jumps more than 15% as post-pandemic bonuses surge

By Lydia Moynihan

June 13, 2022 | 1:58pm | Updated

David Solomon hauled in big bucks in 2021.
Bloomberg via Getty Images

MORE ON:
CEOS

Compensation for chief executives jumped 15.7% last year — driven mainly by huge bonus payouts as corporations recovered from the pandemic, according to a new study.

Relating variables to each other

- We may be interested in the relationship between two variables
- Does GDP per capita relate to Healthy life expectancy?

Country or region	Score	GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
Guatemala	6.436	0.800	1.269	0.746	0.535	0.175	0.078
Yemen	3.380	0.287	1.163	0.463	0.143	0.108	0.077
Netherlands	7.488	1.396	1.522	0.999	0.557	0.322	0.298
...	...	...	...	...	...	...	...
Libya	5.525	1.044	1.303	0.673	0.416	0.133	0.152
Jamaica	5.890	0.831	1.478	0.831	0.490	0.107	0.028
United States	6.892	1.433	1.457	0.874	0.454	0.280	0.128

Covariance and correlation

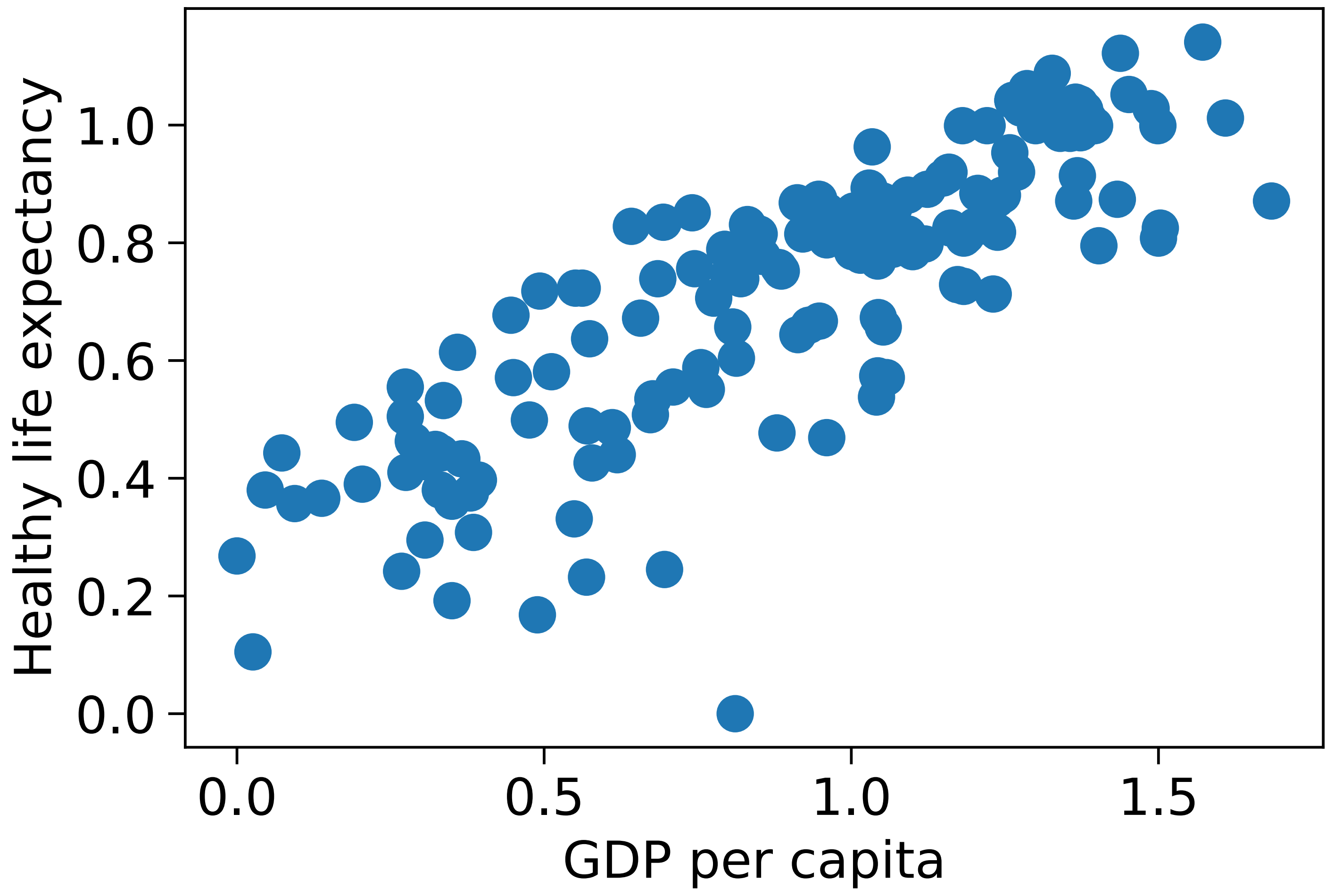
- We have two numerical variables X and Y each with N measurements
- Let's compute the means and SDs of each: $\mu_x, \mu_y, \sigma_x, \sigma_y$
- The covariance σ_{xy} and **Pearson correlation coefficient** ρ_{xy} are given by:

$$\sigma_{xy} = \frac{1}{N} \sum_{n=0}^{N-1} (x^{(n)} - \mu_x)(y^{(n)} - \mu_y)$$

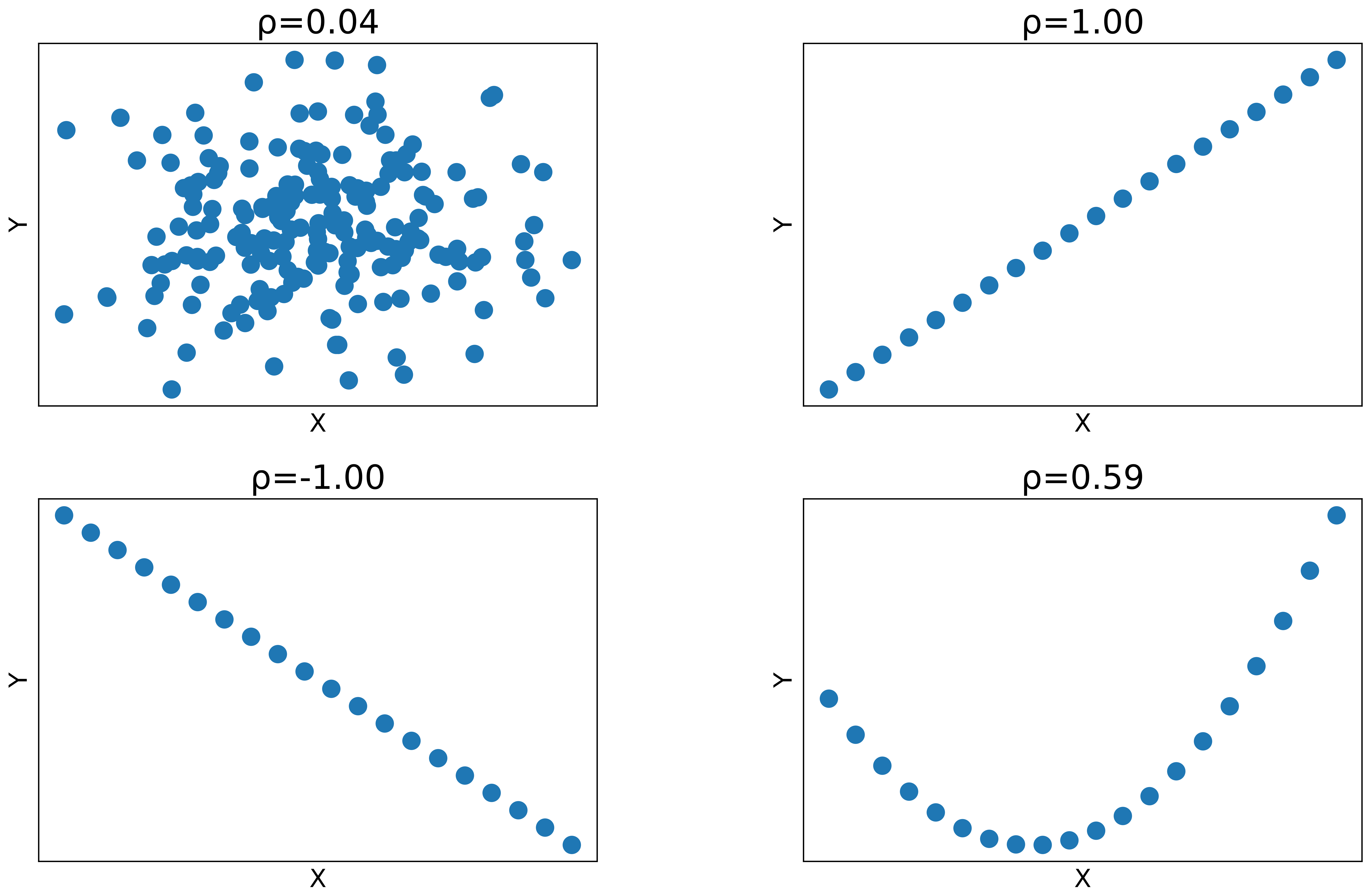
$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Pearson correlation coefficient

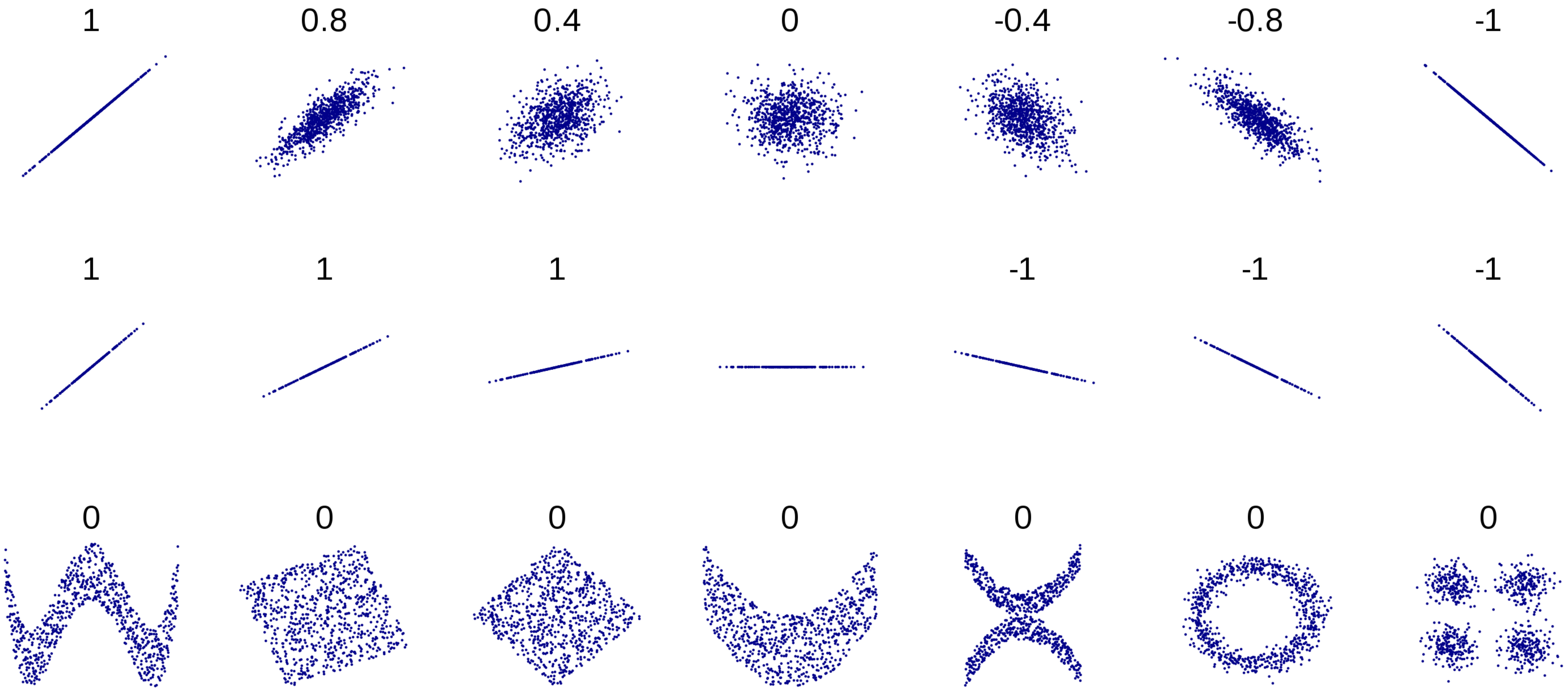
- ρ_{xy} has a value between -1 and 1
- Gives a measure of how linear the relationship between X and Y is
- I.e. the extent to which we can use a line to predict one from the other
- 0.84 for GDP per capita and Healthy life expectancy



Pearson correlation coefficient

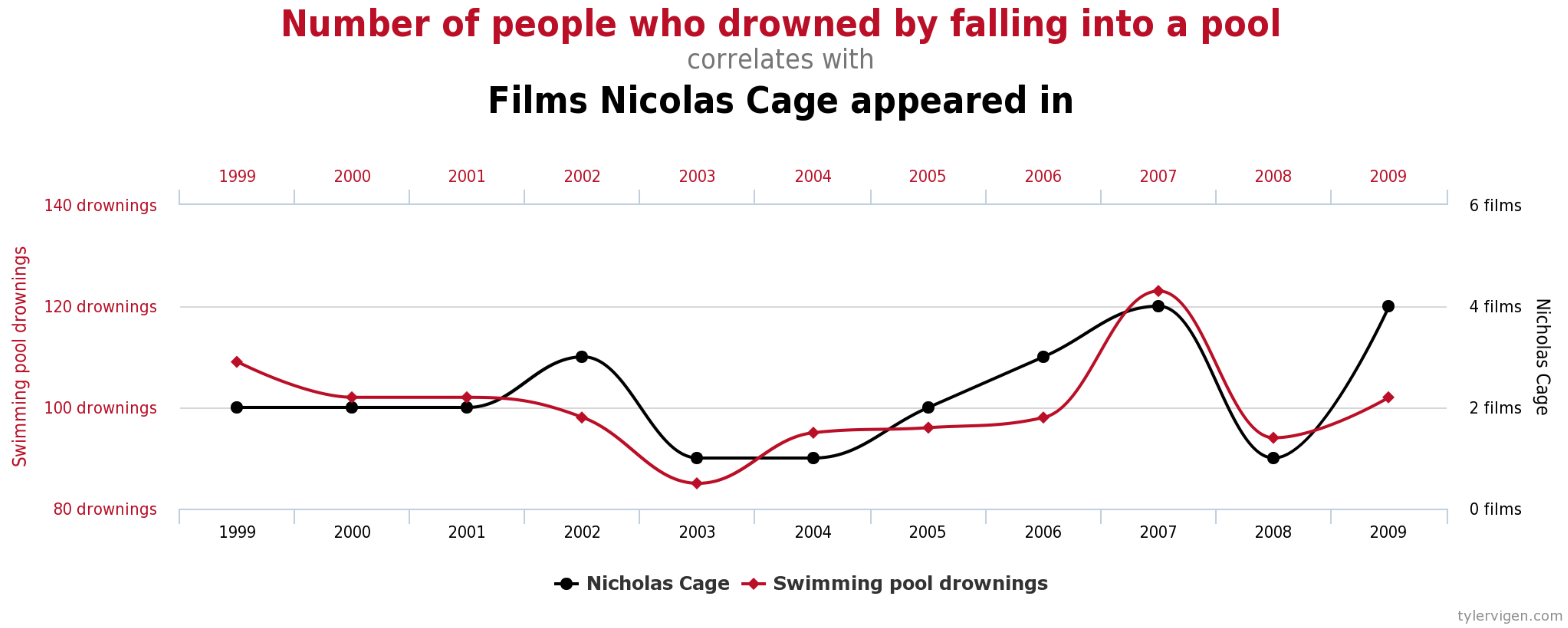


Pearson correlation coefficient



Source: https://en.wikipedia.org/wiki/Correlation#/media/File:Correlation_examples2.svg

Correlation does not imply causation



Rubbish in, rubbish out

If your data is rubbish then anything you extract from it is also rubbish

- You might not have enough data items
- The process for collecting data might be flawed (e.g. biased)
- Measurements might be recorded incorrectly
- The variables chosen might not be useful

Misleading statistics

Can be nefarious, or just stupidity

BBC NEWS

LIVE BBC NEWS CHANNEL

News Front Page
World
UK
England
Northern Ireland
Scotland
Wales
Business
Politics
Health
Education
Science & Environment
Technology
Entertainment
Also in the news

Video and Audio

Have Your Say
Magazine
In Pictures
Country Profiles
Special Reports

RELATED BBC SITES
SPORT
WEATHER
CBBC NEWSROUND

Last Updated: Wednesday, 17 January 2007, 02:45 GMT

E-mail this to a friend   Printable version

Colgate warned over '80%' boast

The maker of Colgate toothpaste has been warned not to repeat its famous advertising claim that "more than 80% of dentists recommend Colgate".

Colgate's claim on posters was "misleading"

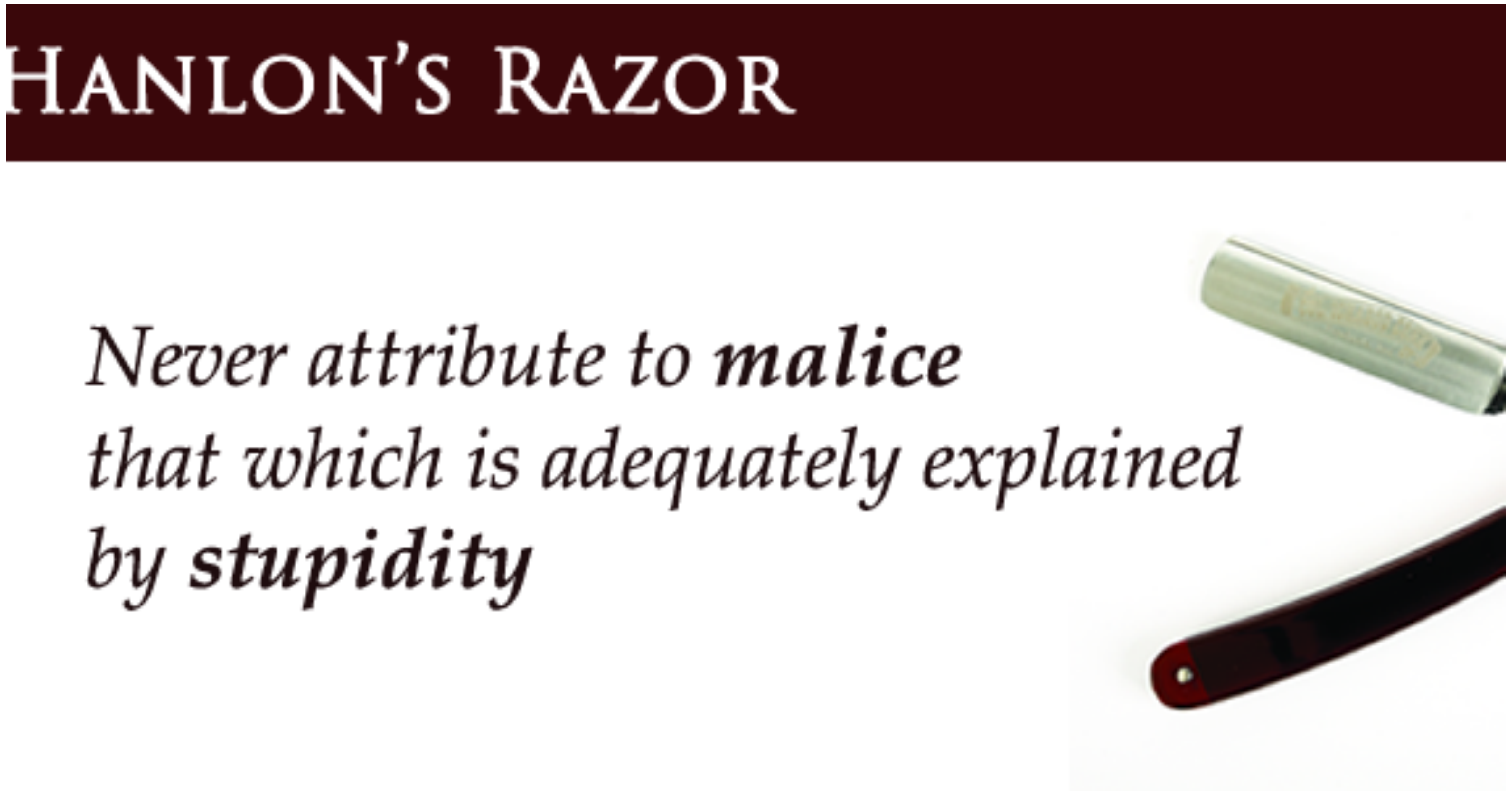
The Advertising Standards Authority concluded the claim on Colgate posters was "misleading" after investigating the phone survey behind the boast.

It found the dentists surveyed were allowed to name more than one brand.

But the ASA said the advertising claim implied 80% of dentists recommended Colgate to the exclusion of its rivals.

In fact, the ASA's inquiry found another competitor's brand was recommended almost as much as Colgate was by those dentists who were surveyed.

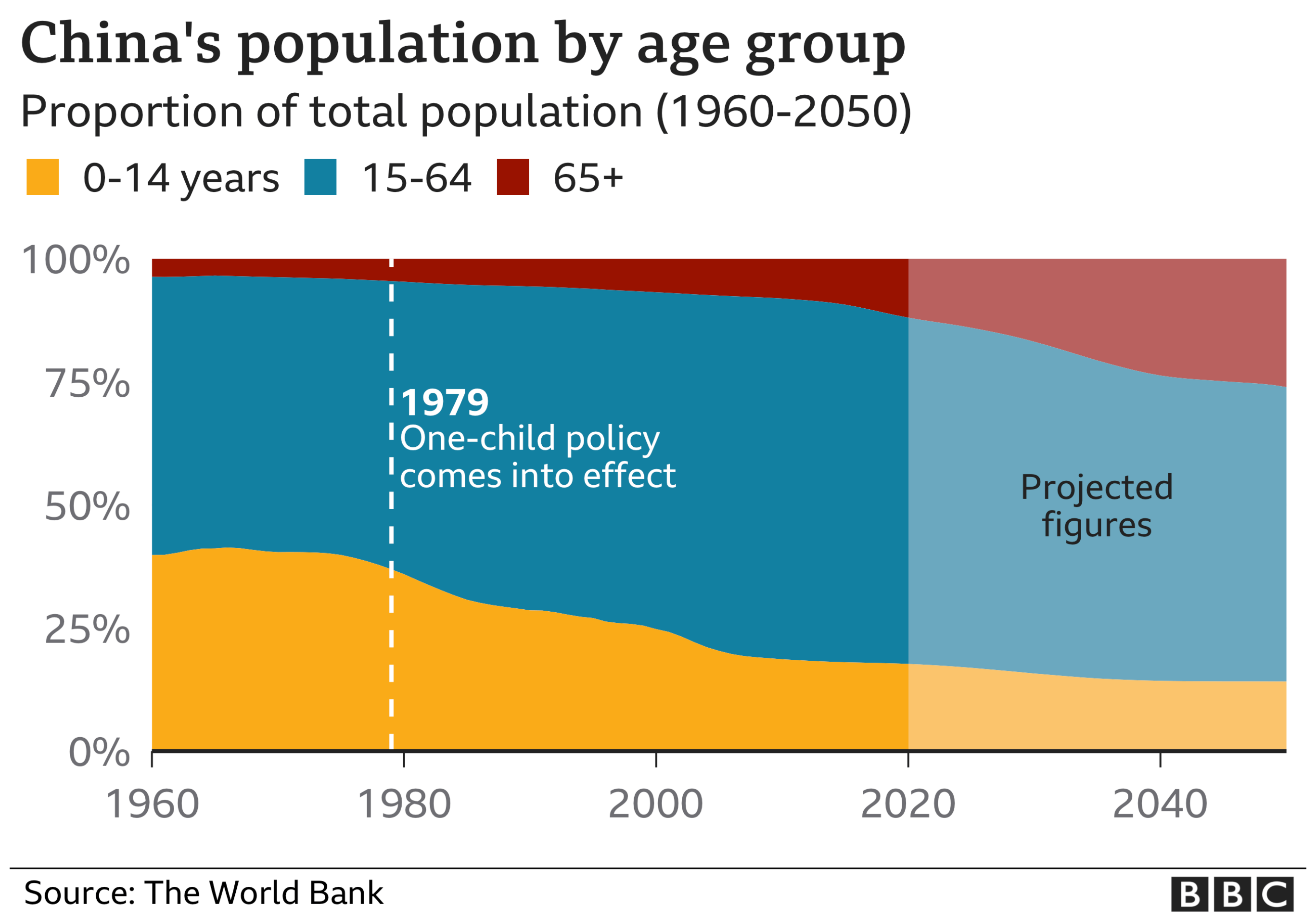
It added the survey "did not make clear the poll was on behalf of Colgate".

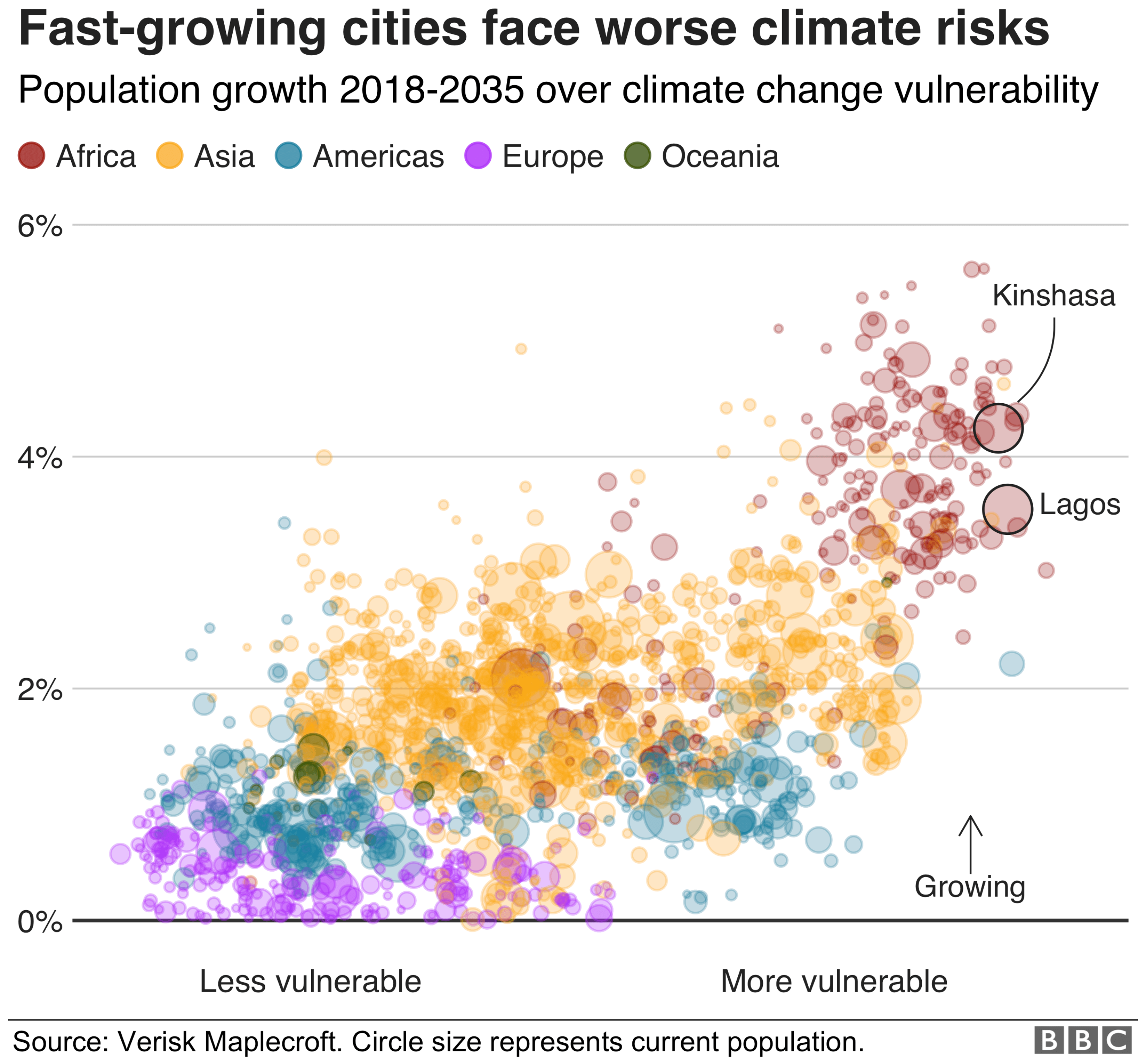


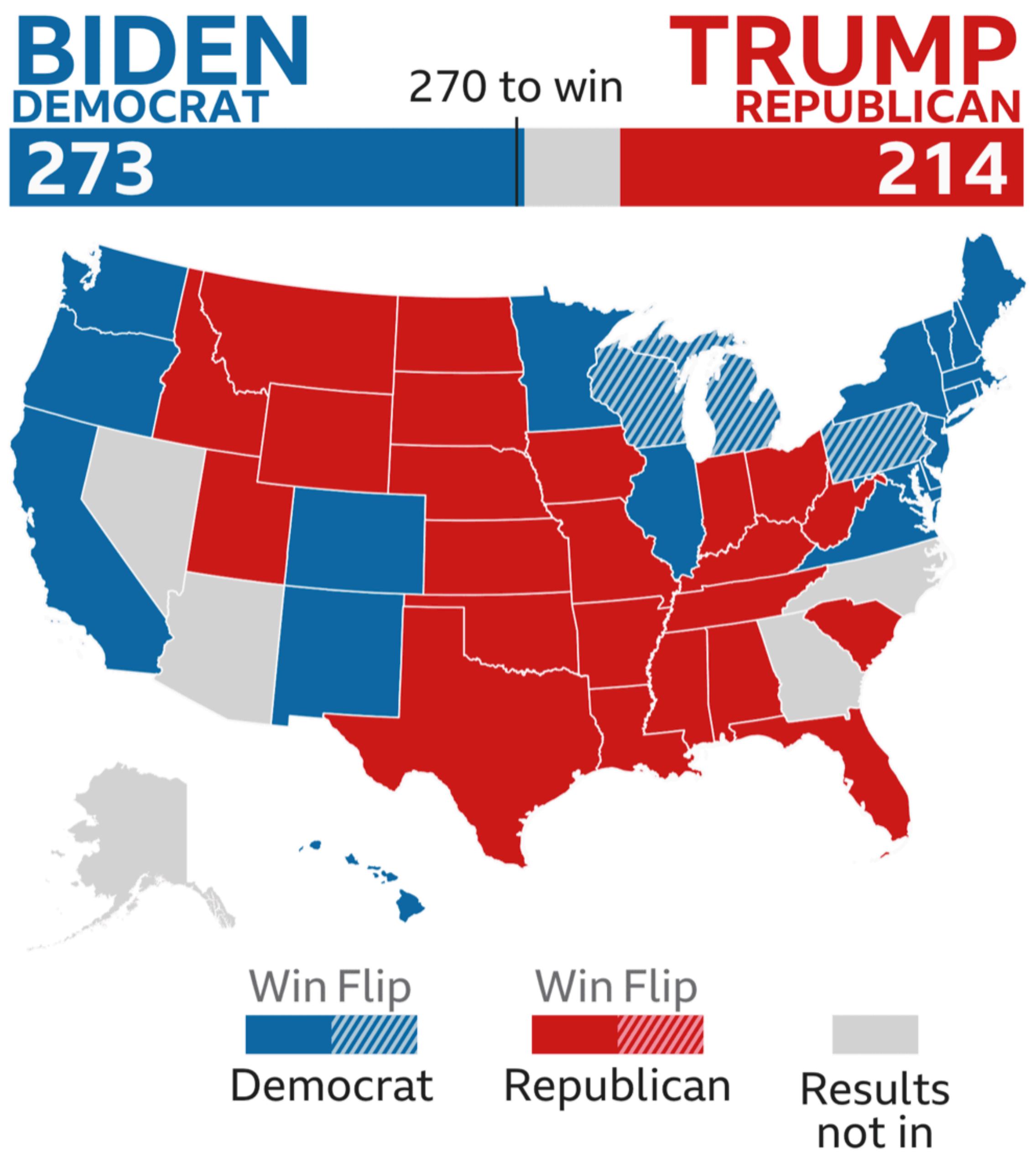
Visualising Data

Visualising data for presentation

- Conveying information as **simply**, and **clearly** as possible
- It is an art form, combining data analysis with graphics design



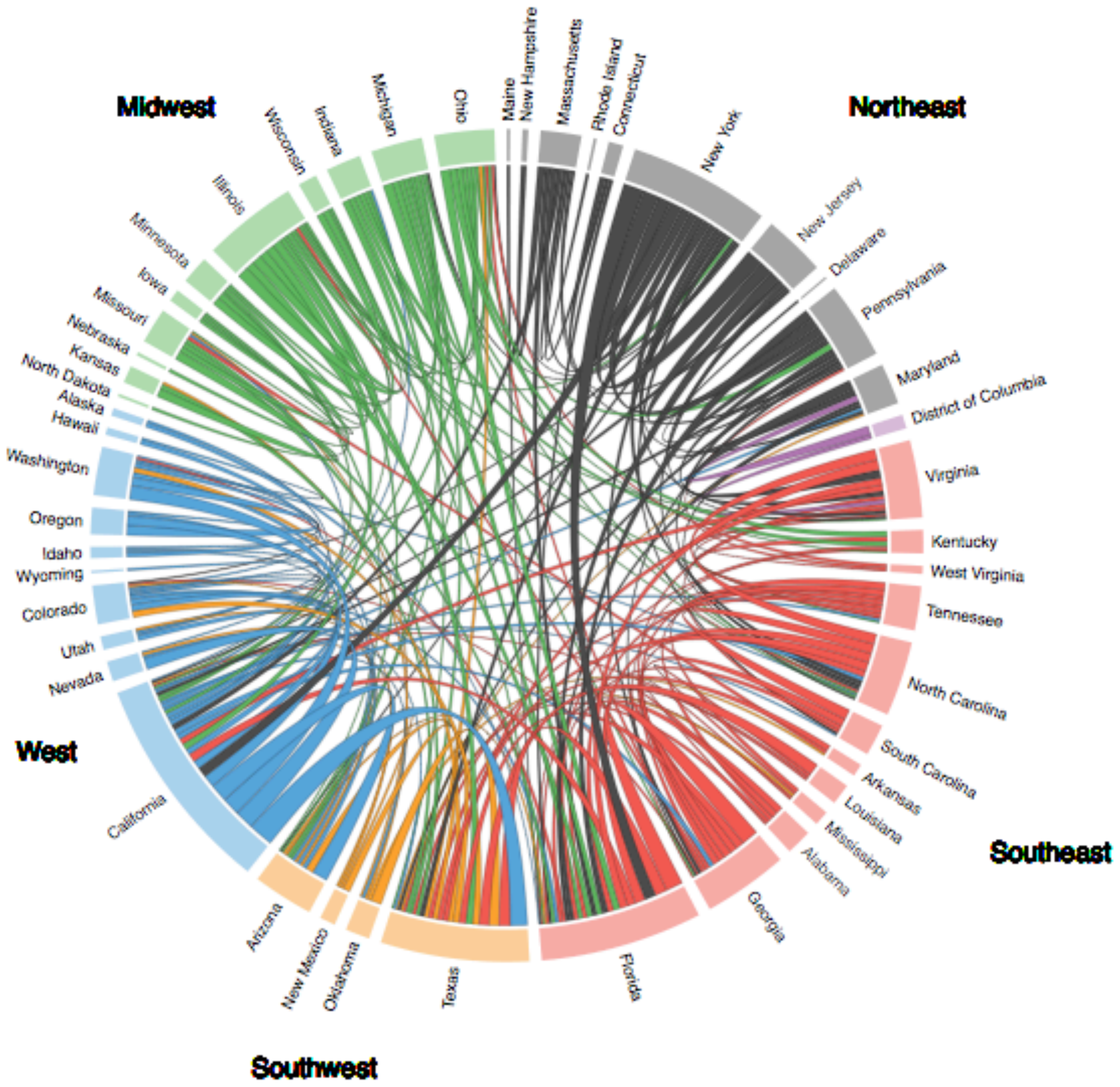


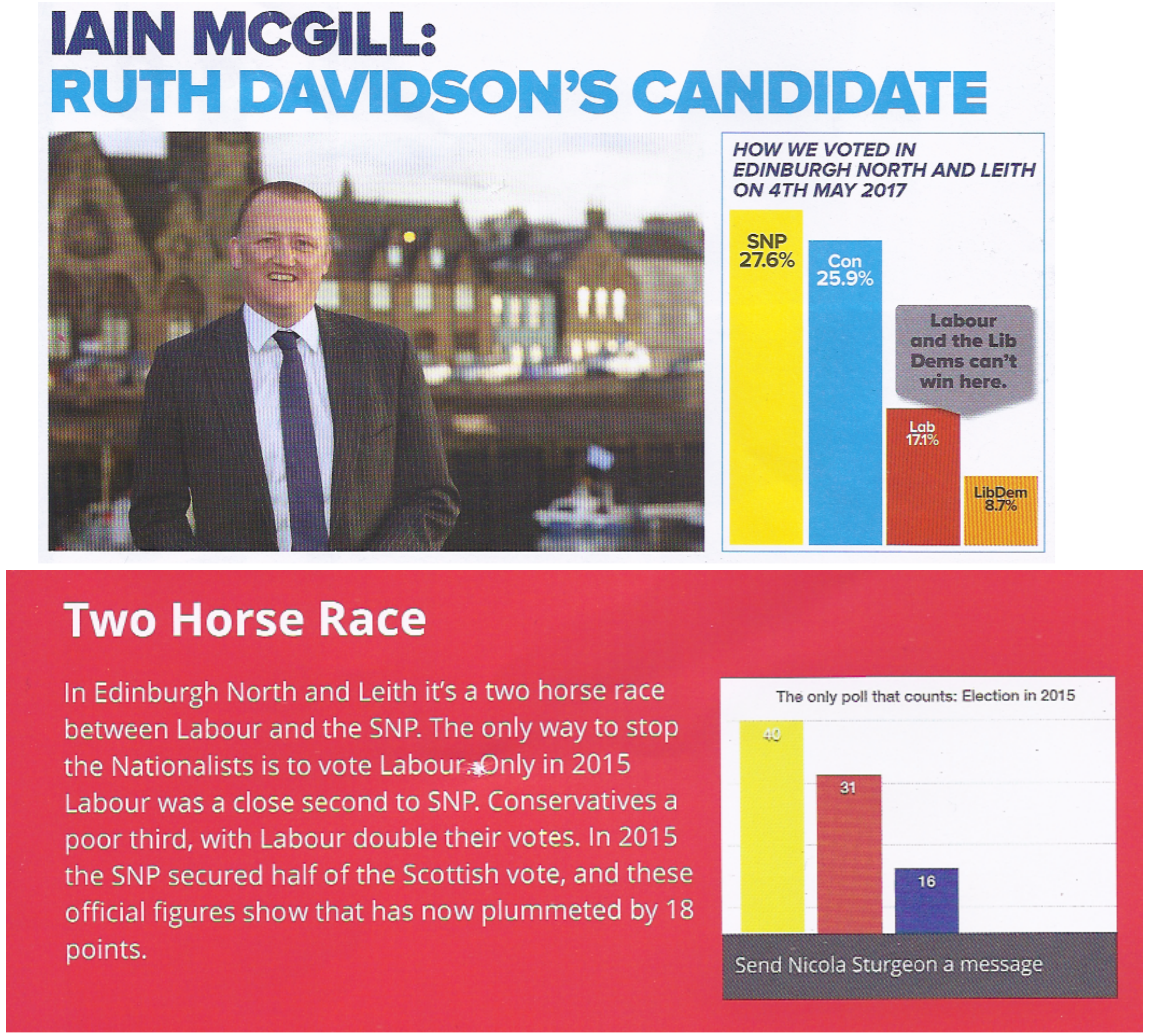


Source: BBC

Visualising data for presentation

Can be done badly e.g. overcomplicated or misleading



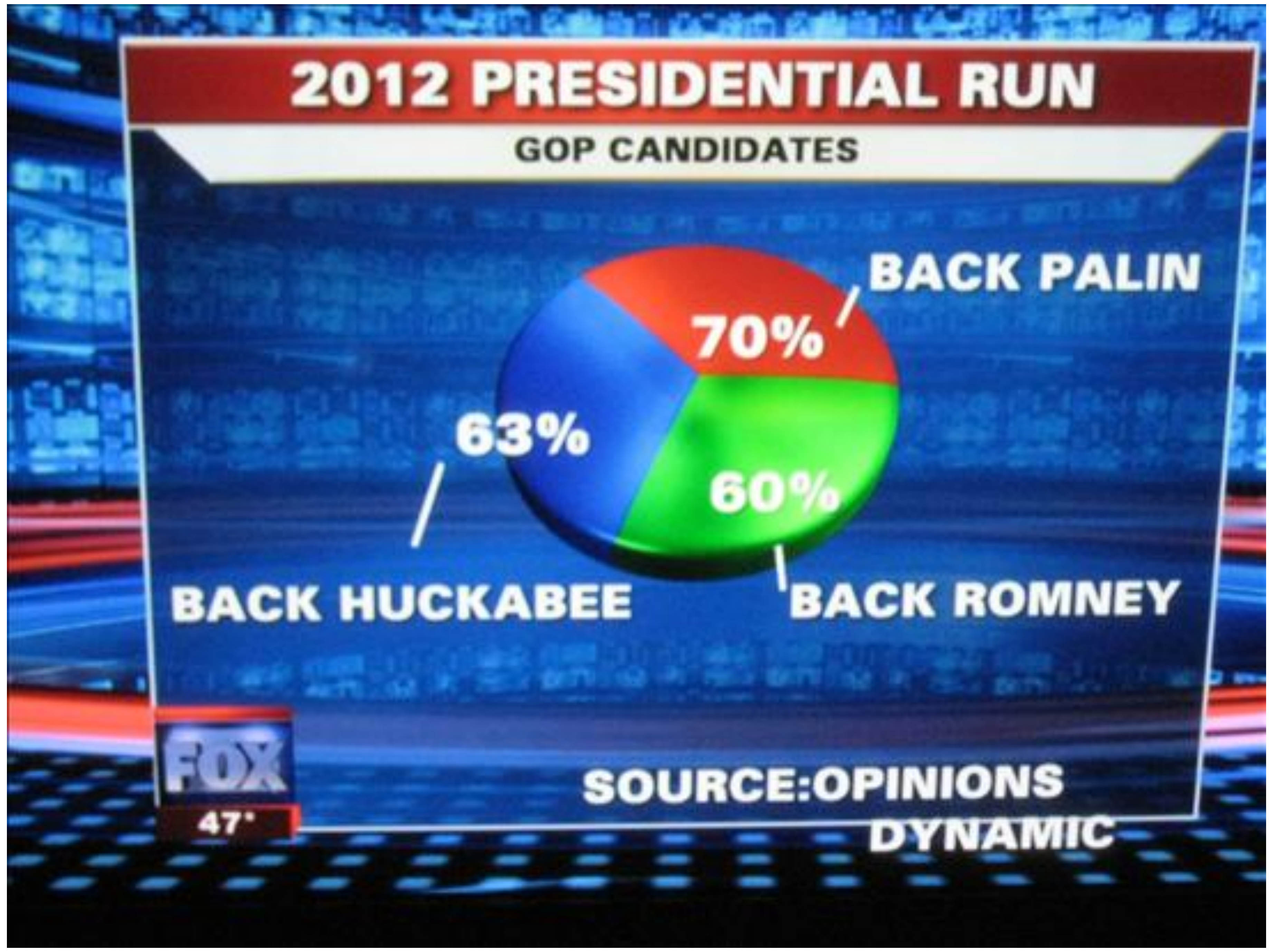


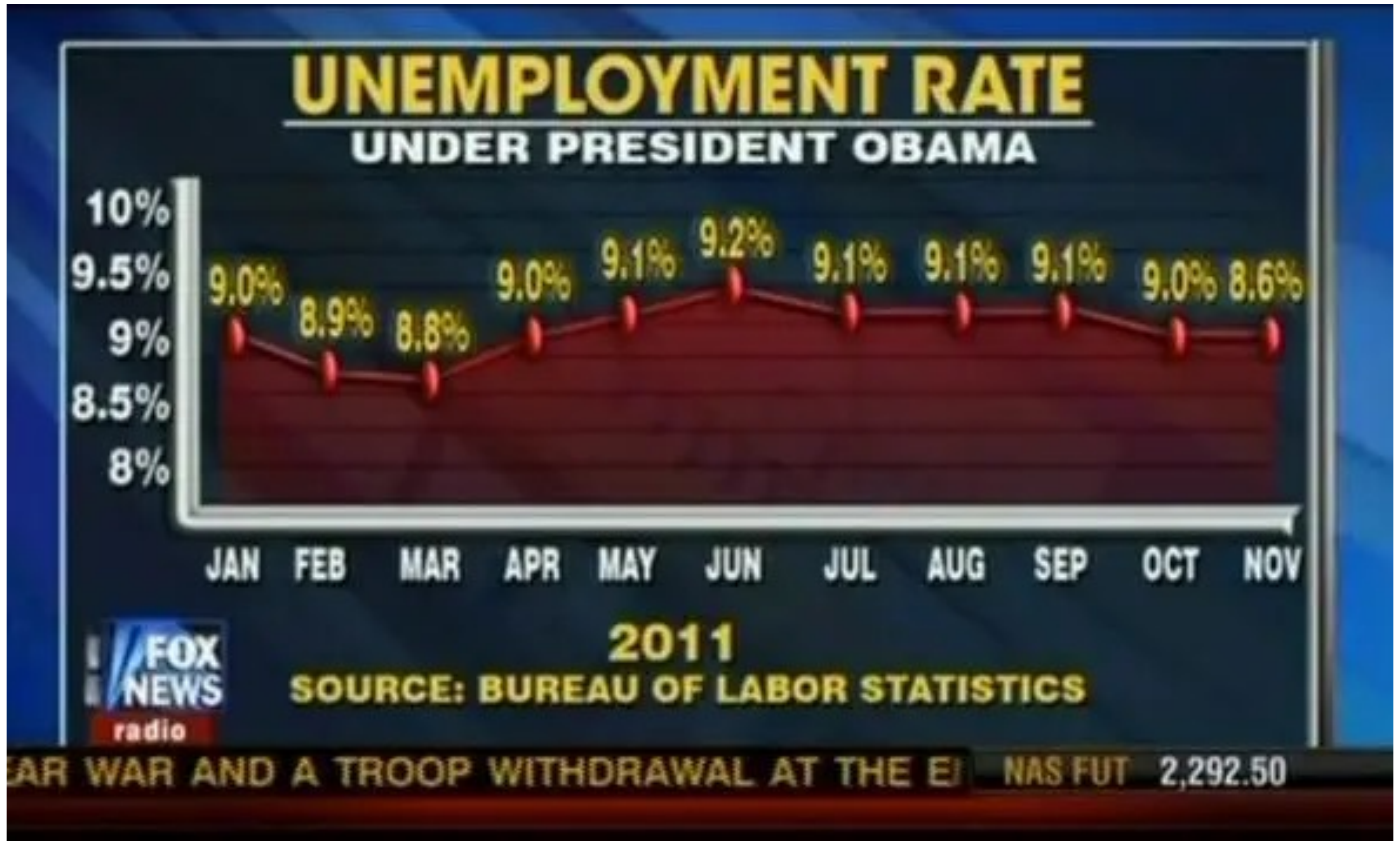
Source: https://junkcharts.typepad.com/junk_charts/2014/01/visualizing-movements-of-people.html

Source: https://www.broughtonspurtle.org.uk/news/dodgy-election-bar-charts-4

Visualising data for presentation

Or can just be completely wrong

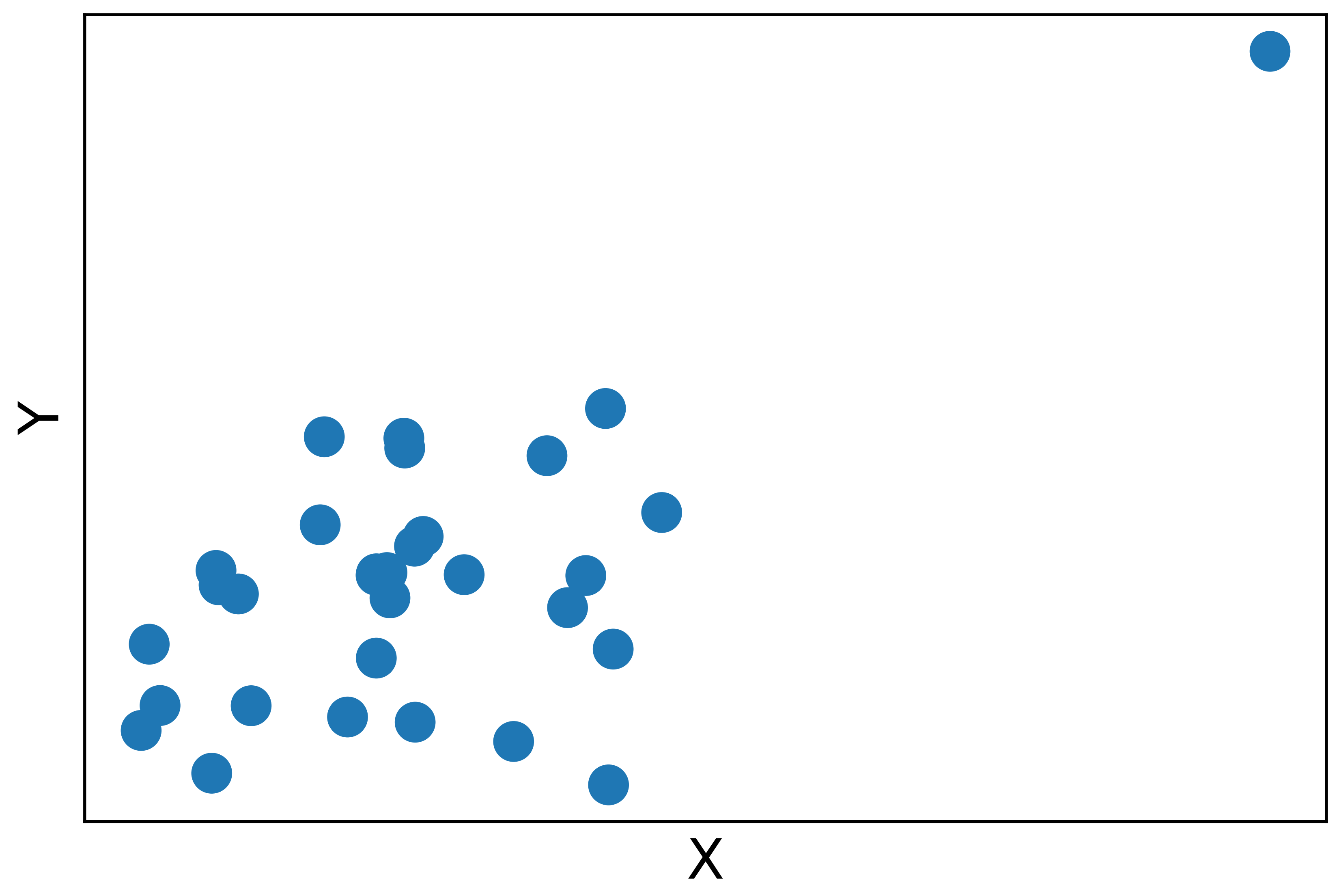


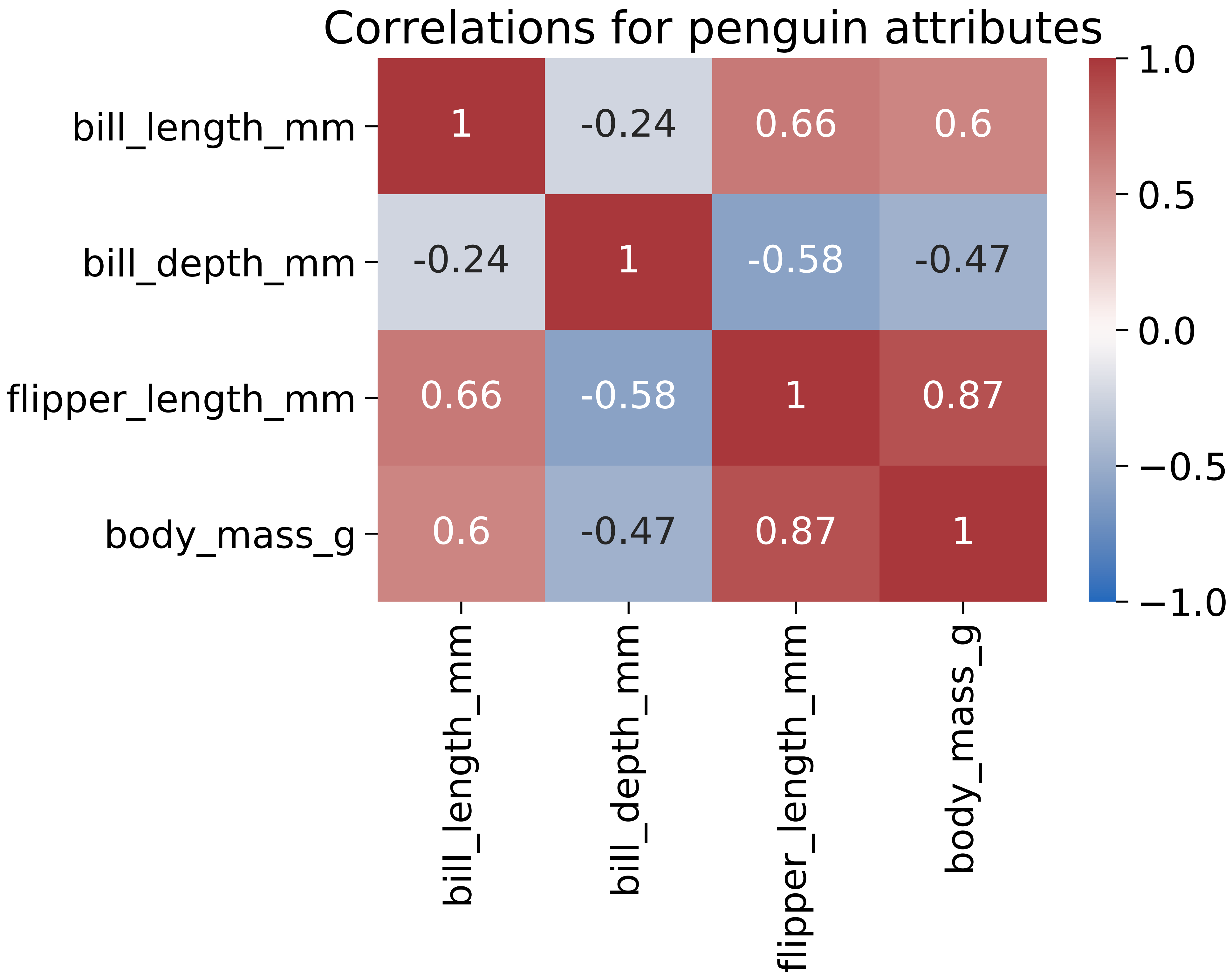


Source: https://www.businessinsider.com/fox-news-charts-tricks-data-2012-11?r=US&IR=T

Visualising data for exploration

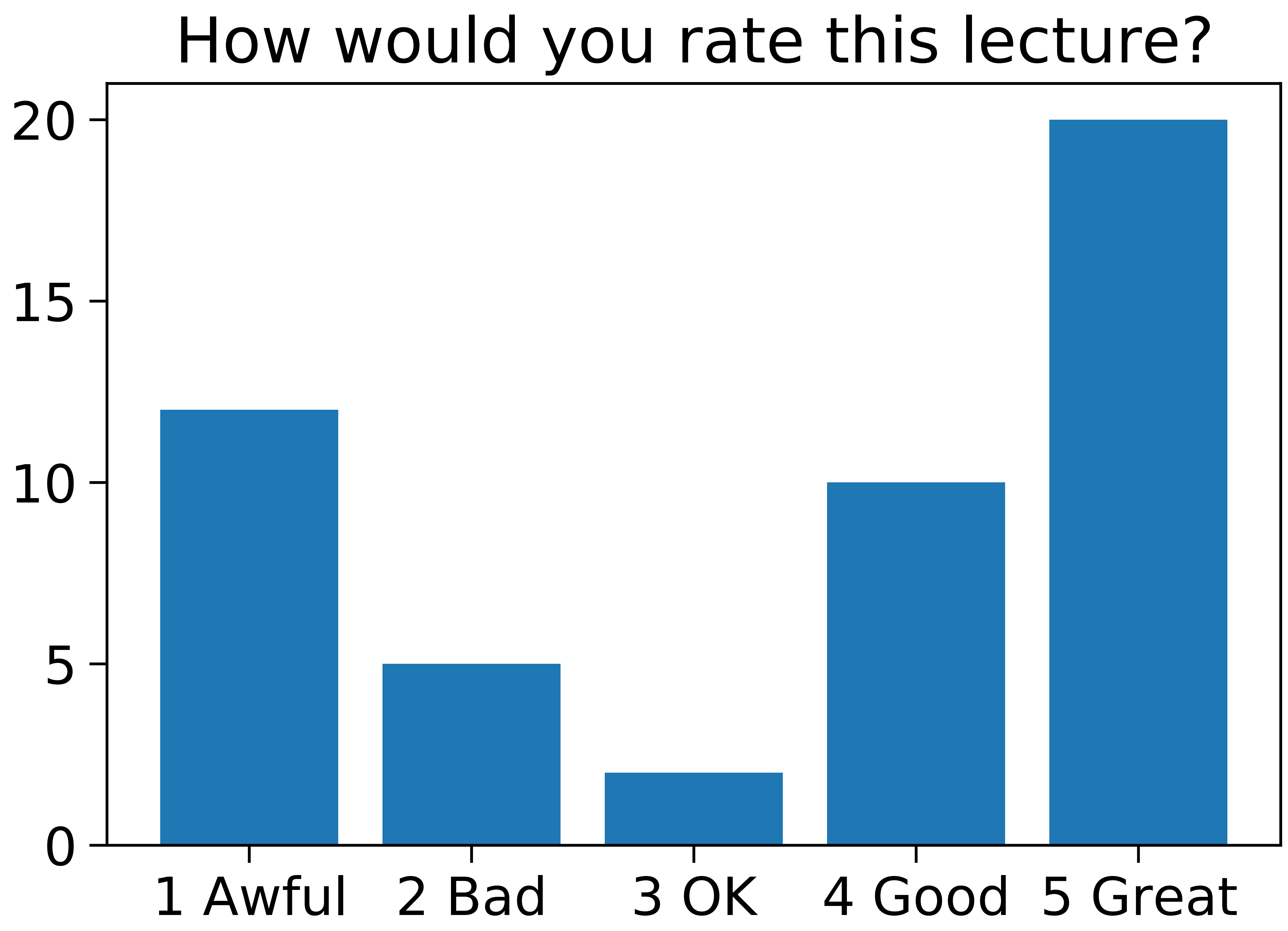
- Finding patterns, spotting outliers and errors, identifying important variables
- Deciding which machine learning method to apply





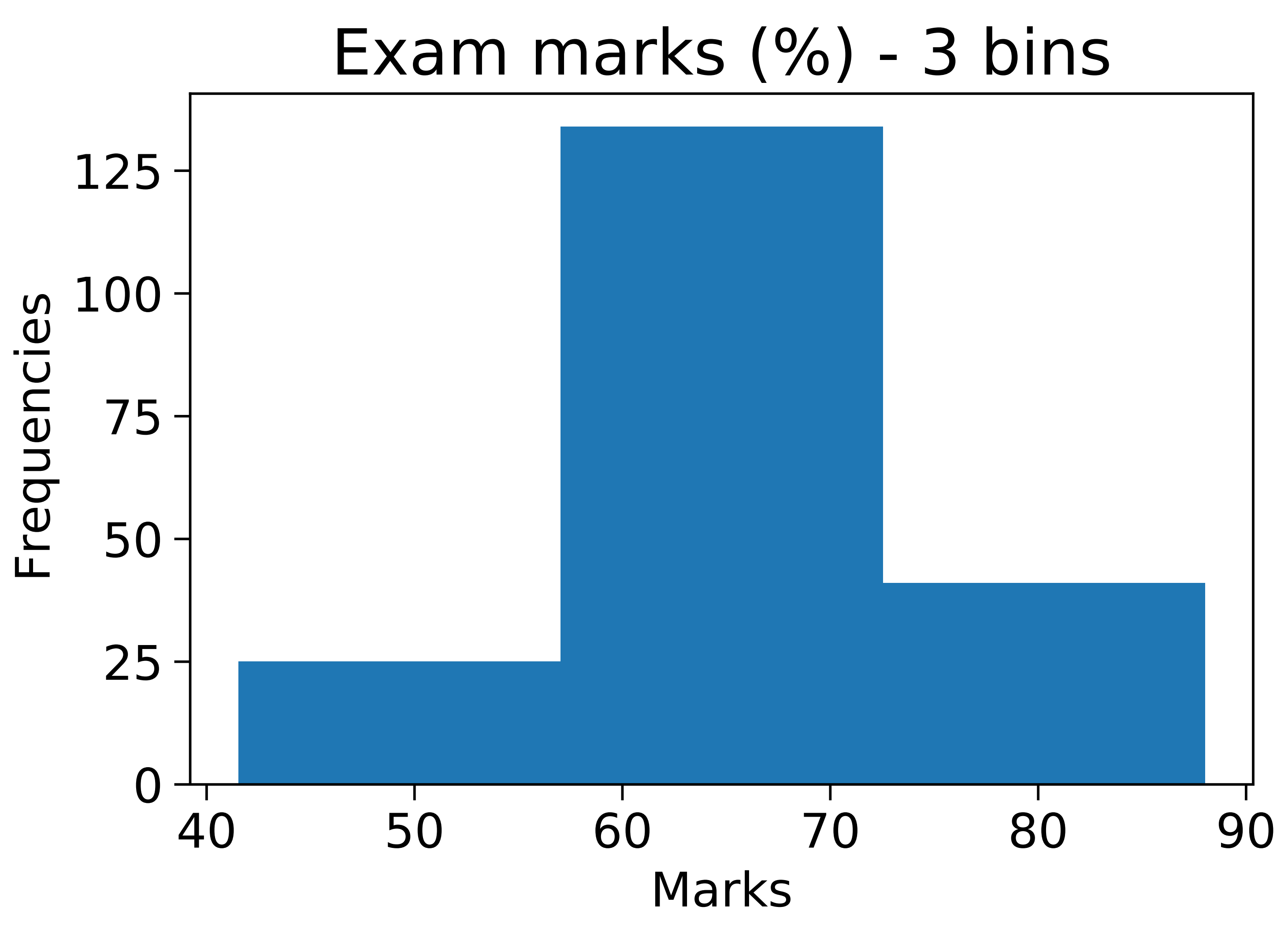
Bar plots

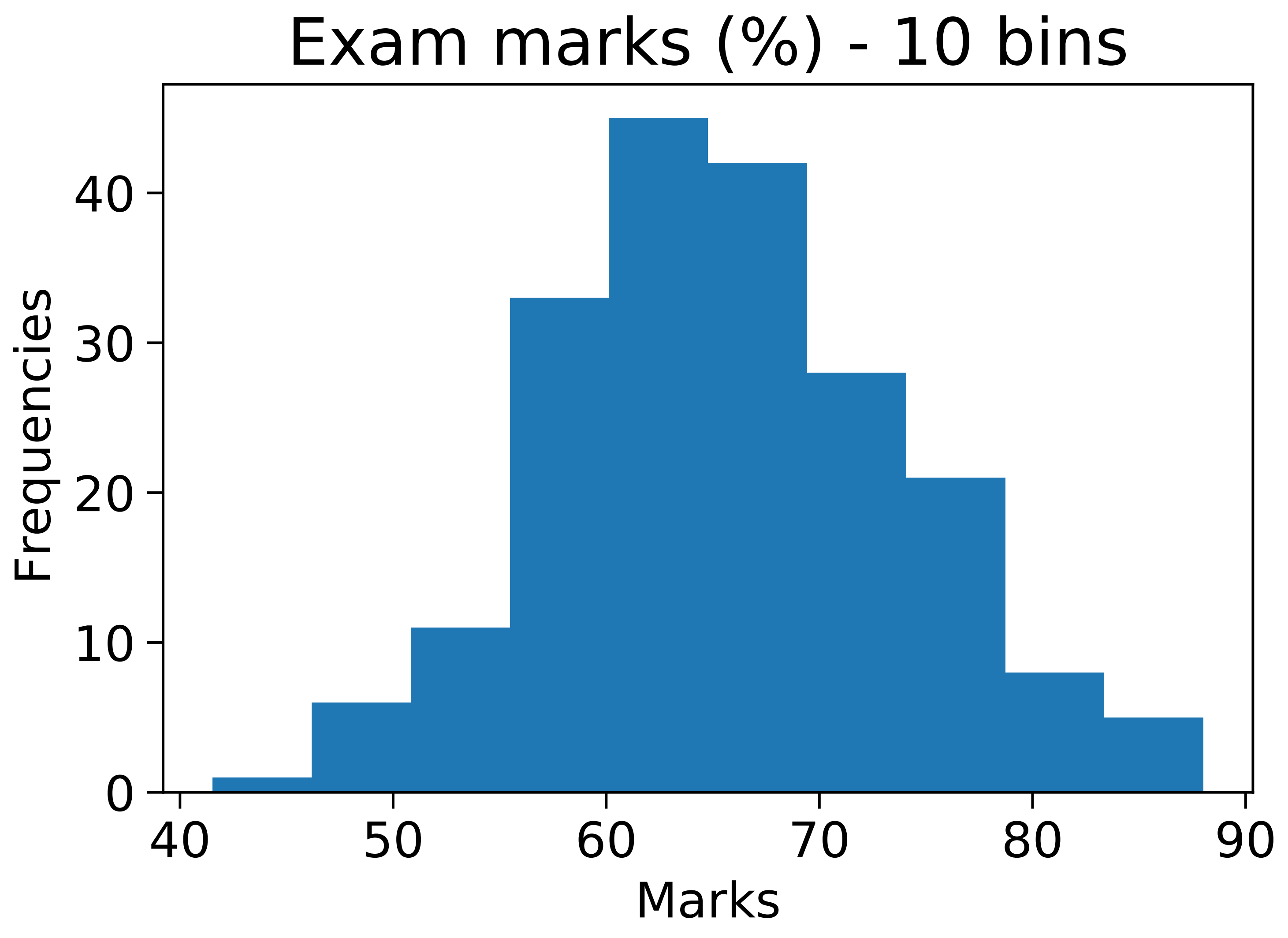
- Good for visualising categorical variables
- If the variable is ordinal then make sure that the columns are in order

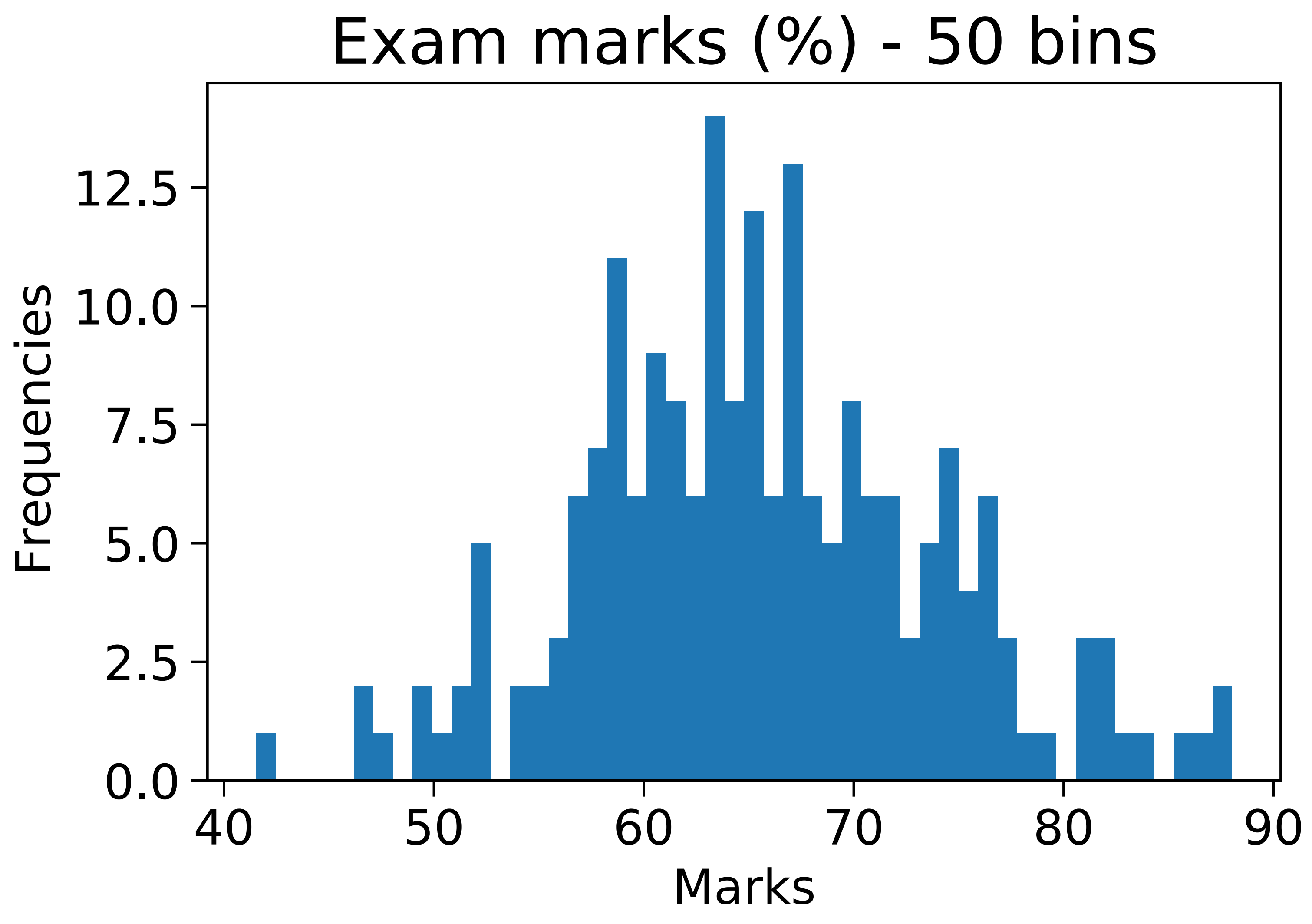


Histograms

- Sorts measurements for numerical variables into equal sized bins
- The number of bins (and/or bin width) may need tweaking depending on use



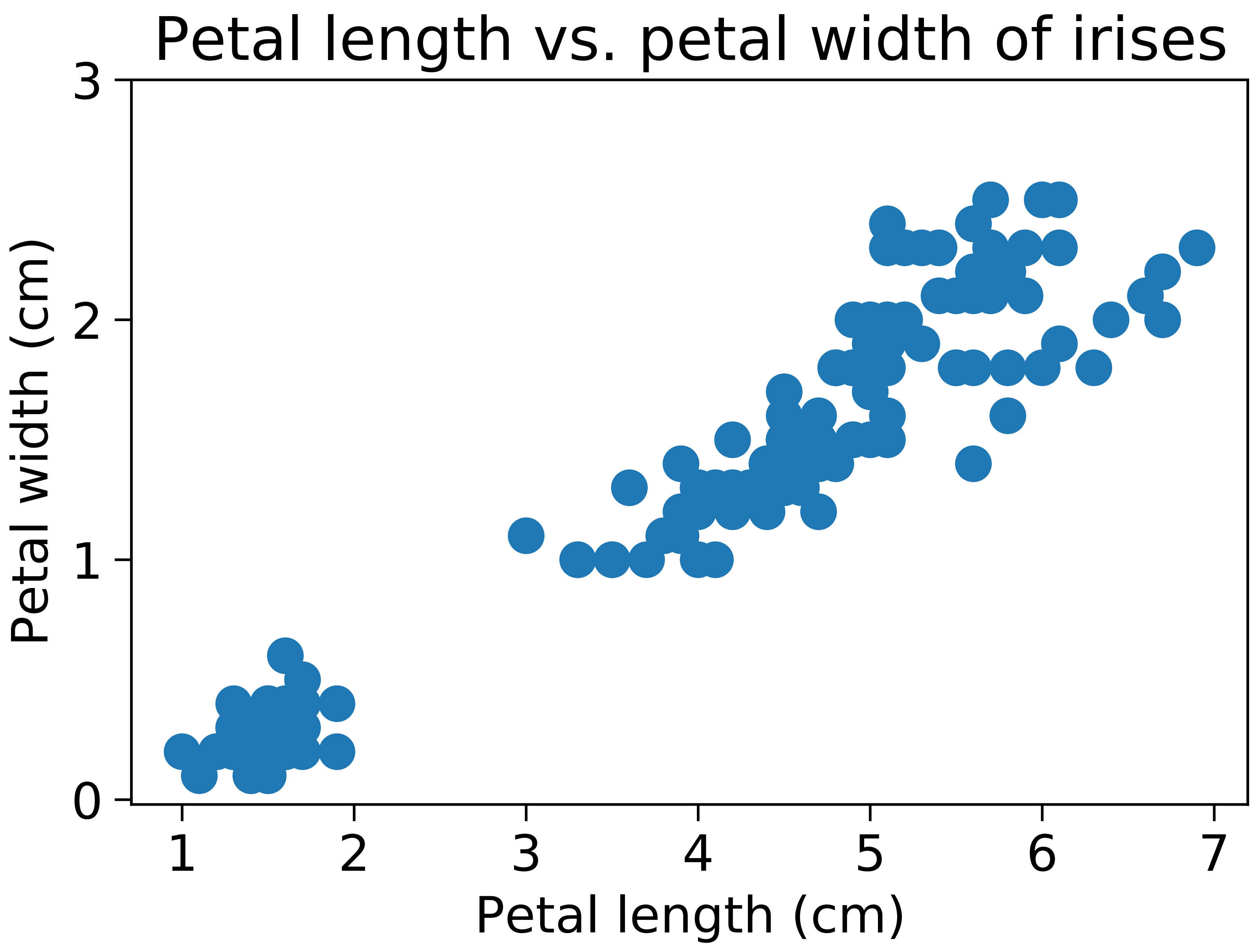


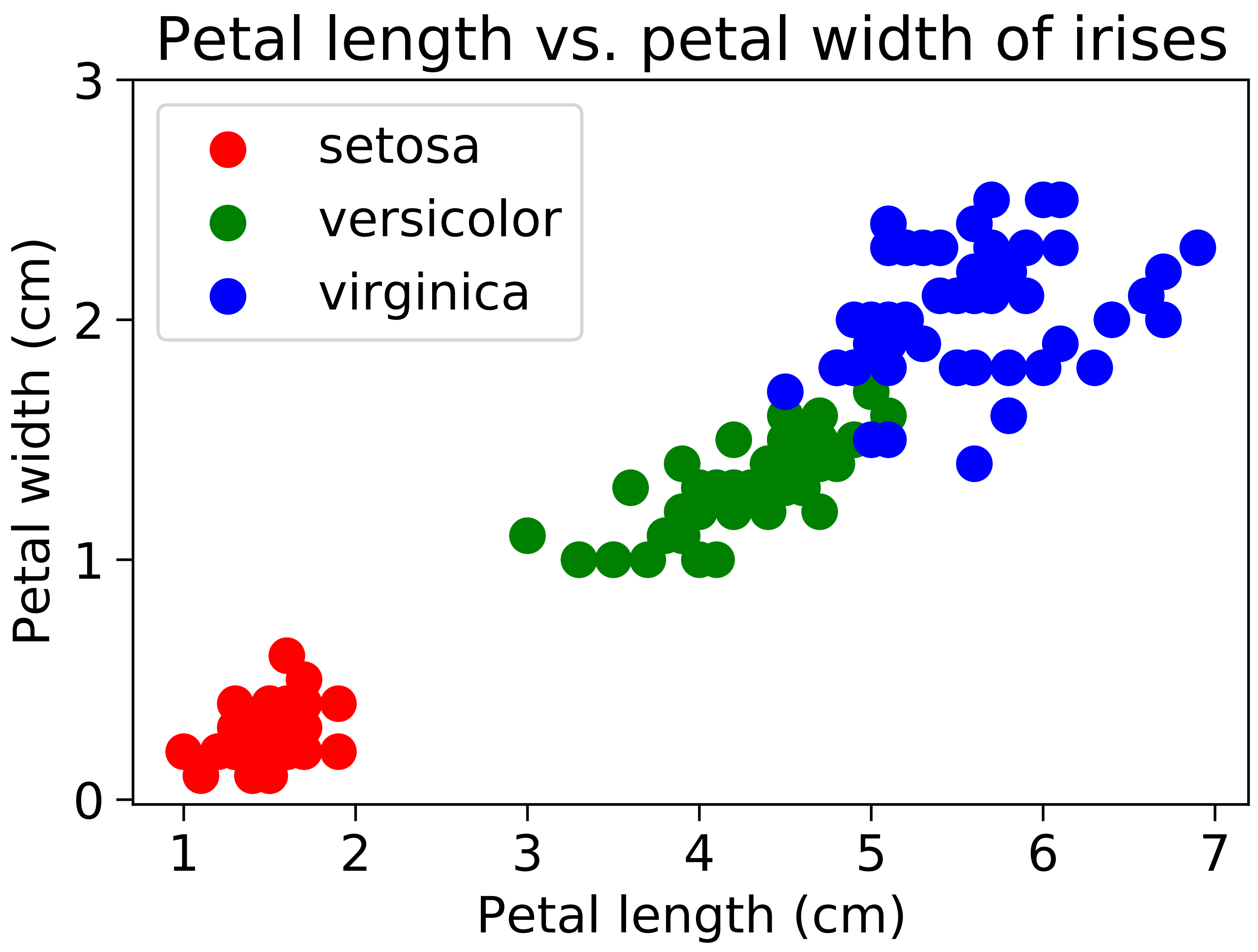


Strange y ticks on this plot.
This can also be tweaked!

Scatter plots in 2D

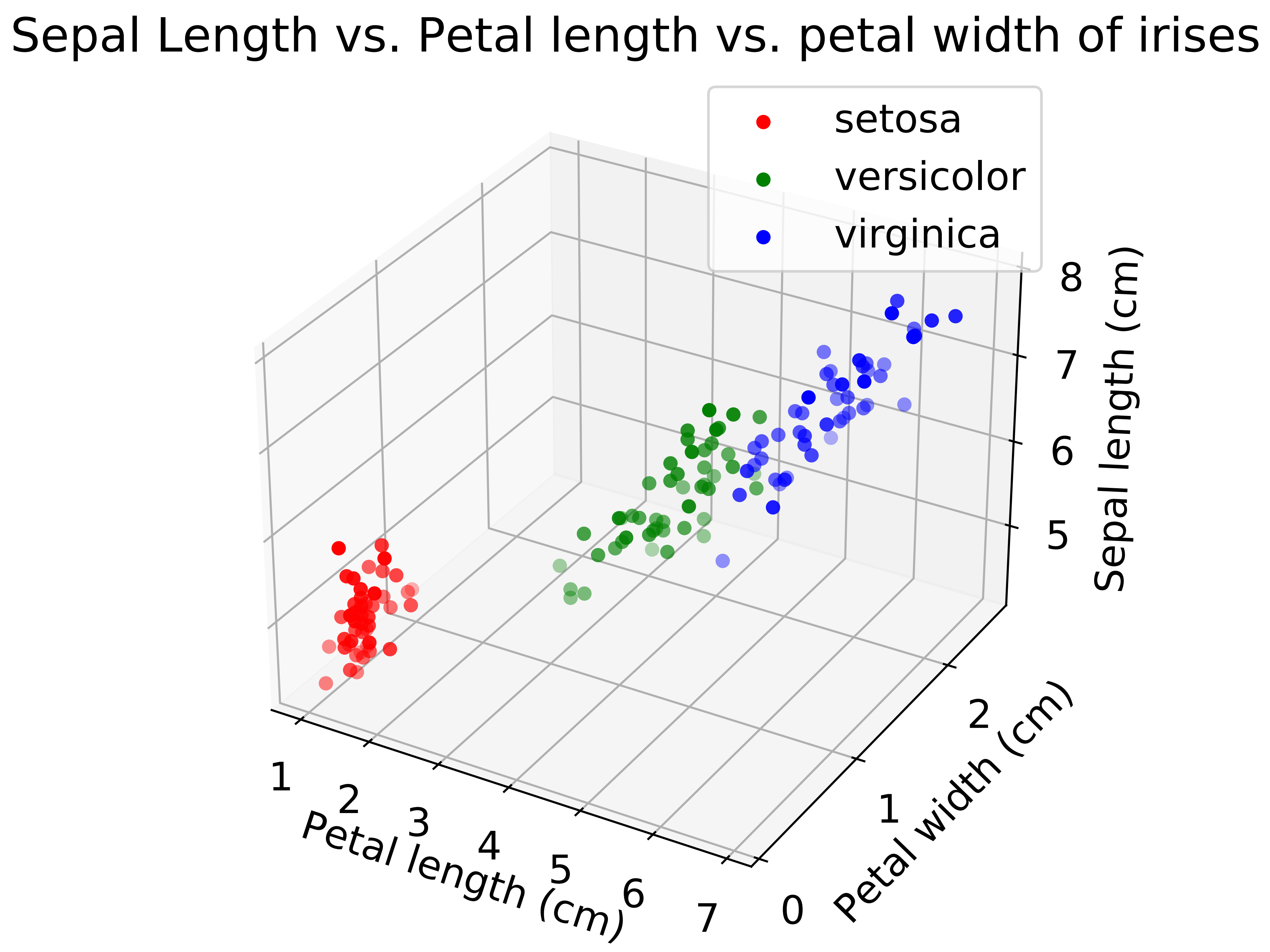
- Each point corresponds to a data item
- The x, y values for that point are measurements of two numerical variables
- We can also distinguish points by category e.g. by using different colours





Scatter plots in 3D

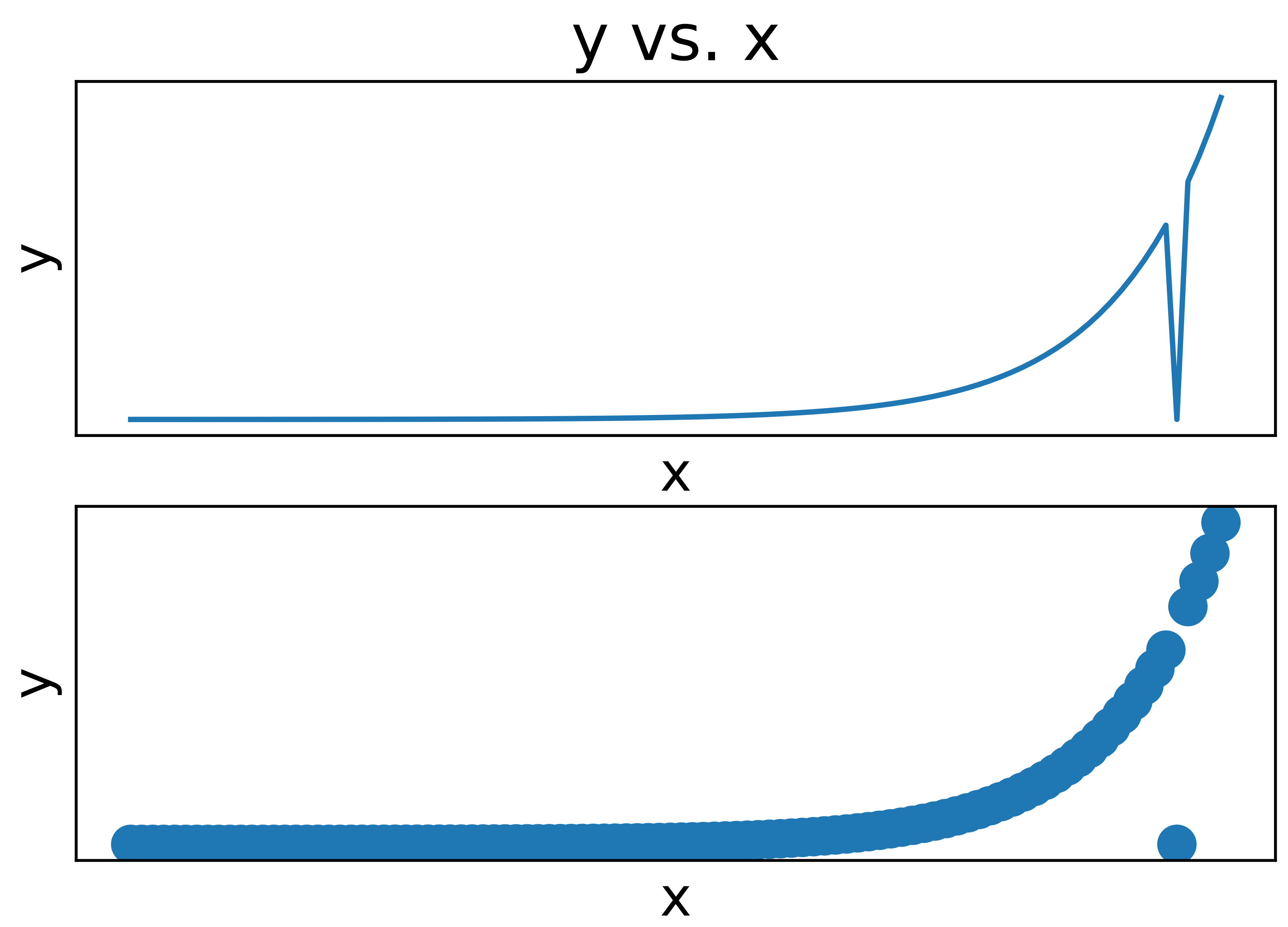
- We can have x, y, z values to show three measurements per point
- But beware: we can't see space properly as its only a 2D projection :(

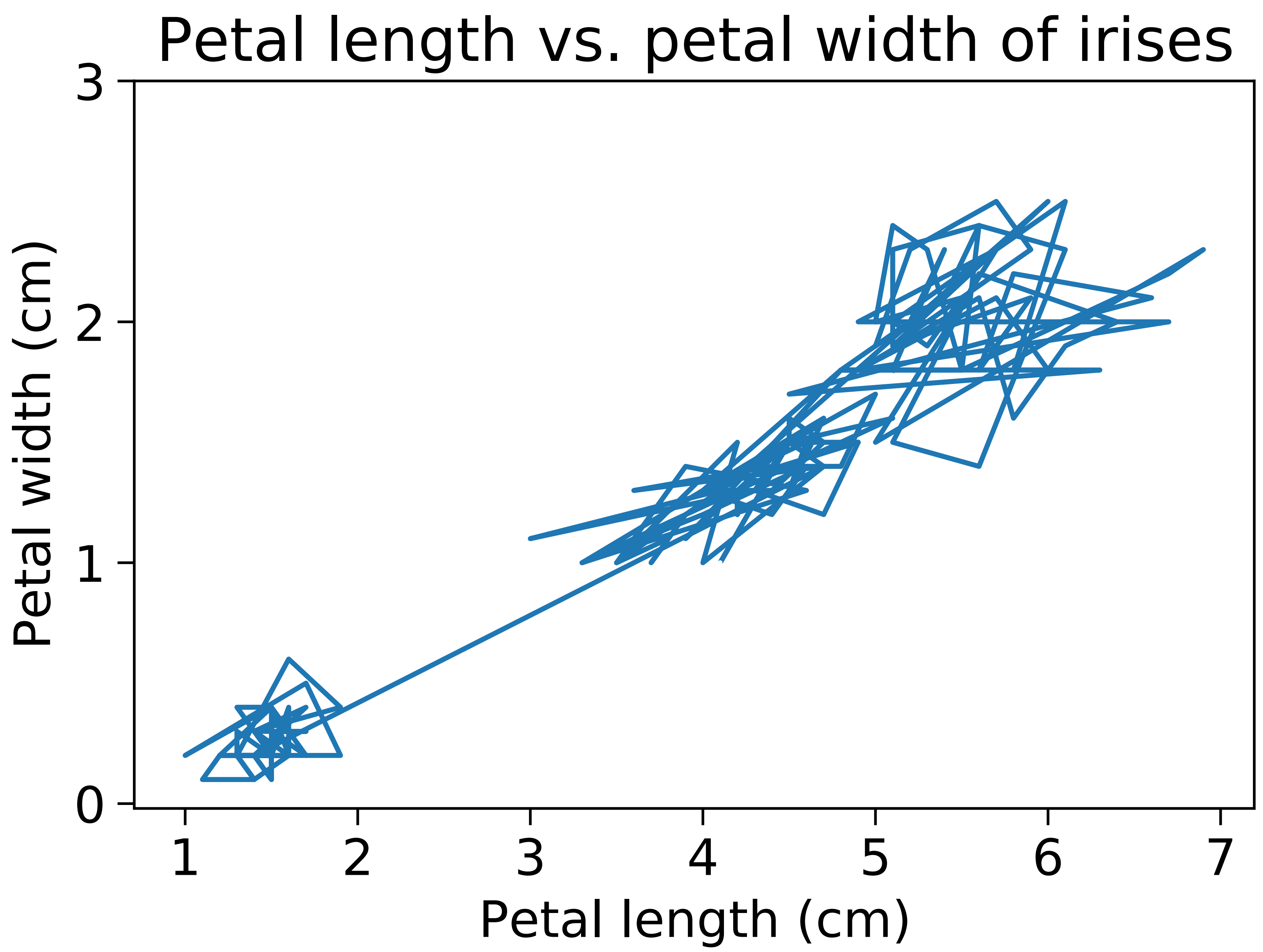


I avoid 3D plots when I can!

Line plots

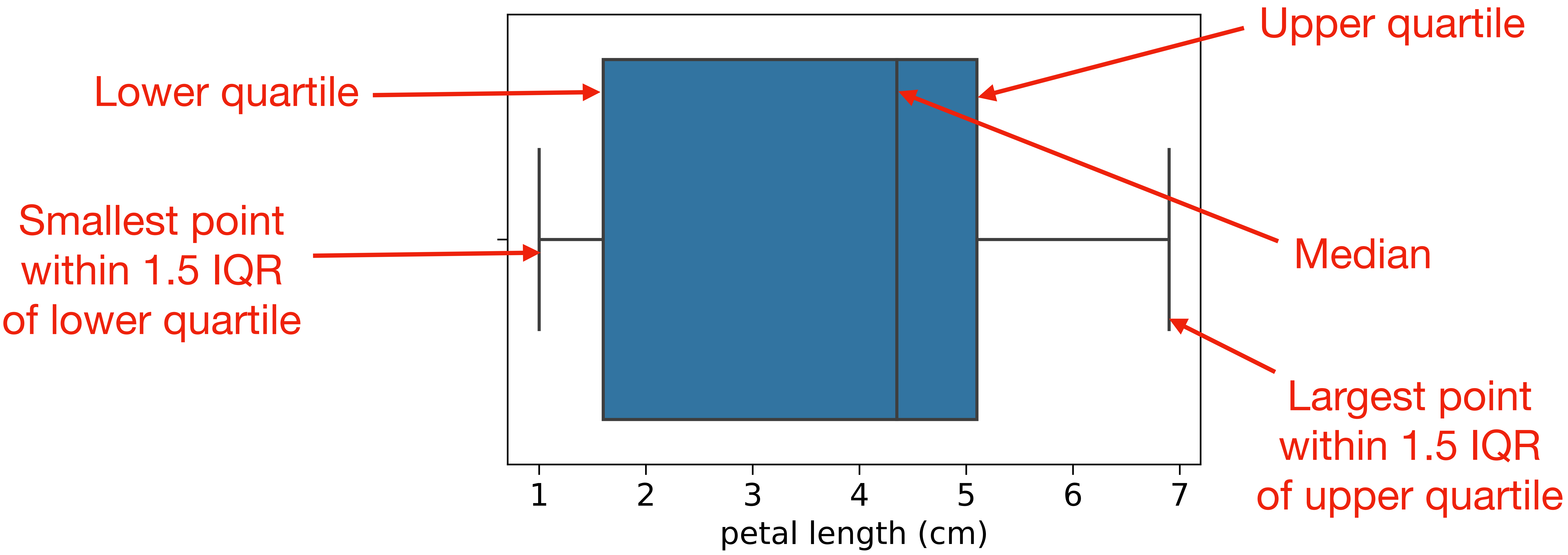
- Can be useful for interpolation
- But can also depict a functional relationship that doesn't exist





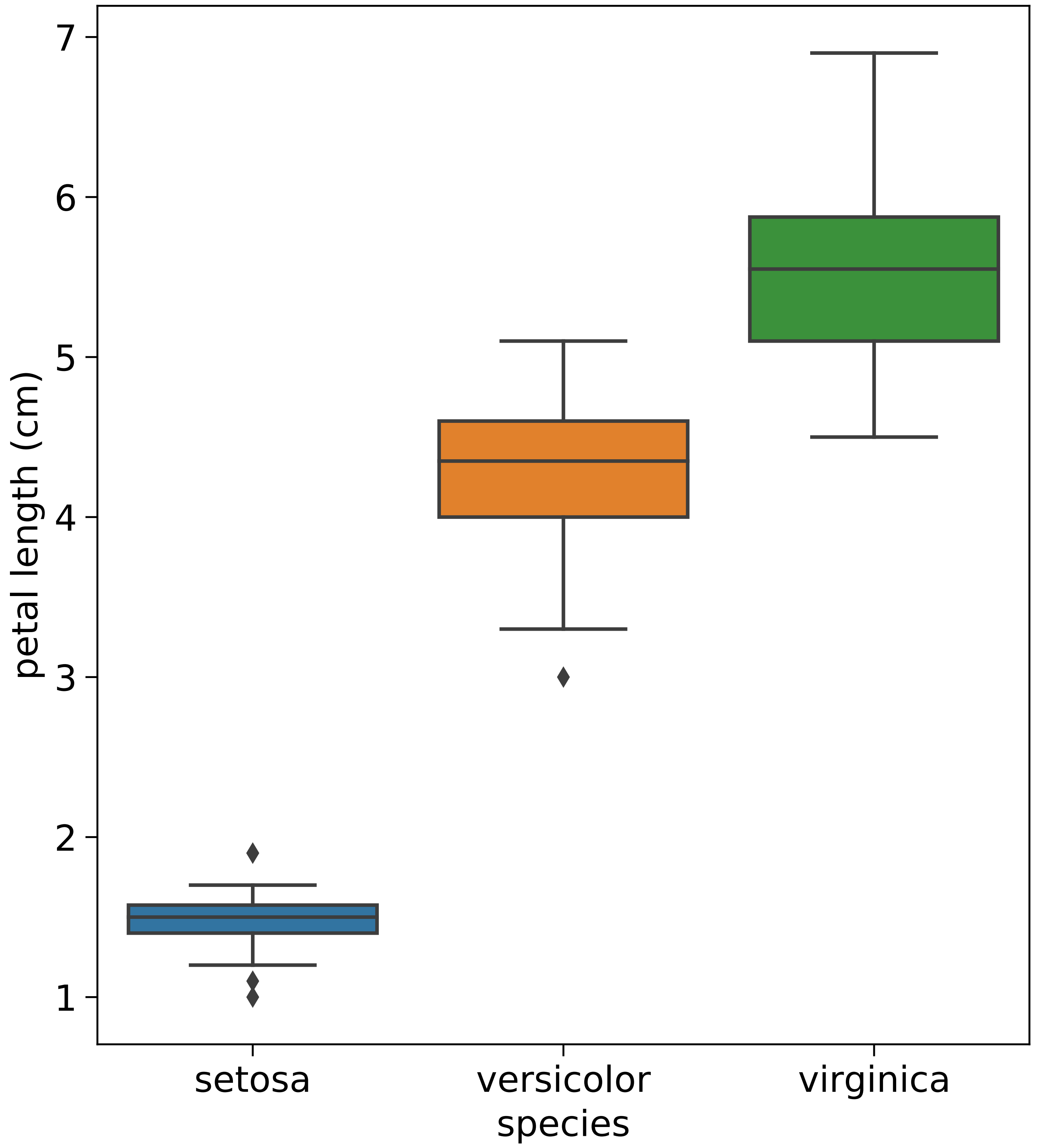
Box plots

- Shows 5 key statistics of a variable, each being an actual measurement
- Interquartile range (IQR) = upper quartile - lower quartile



Box plots

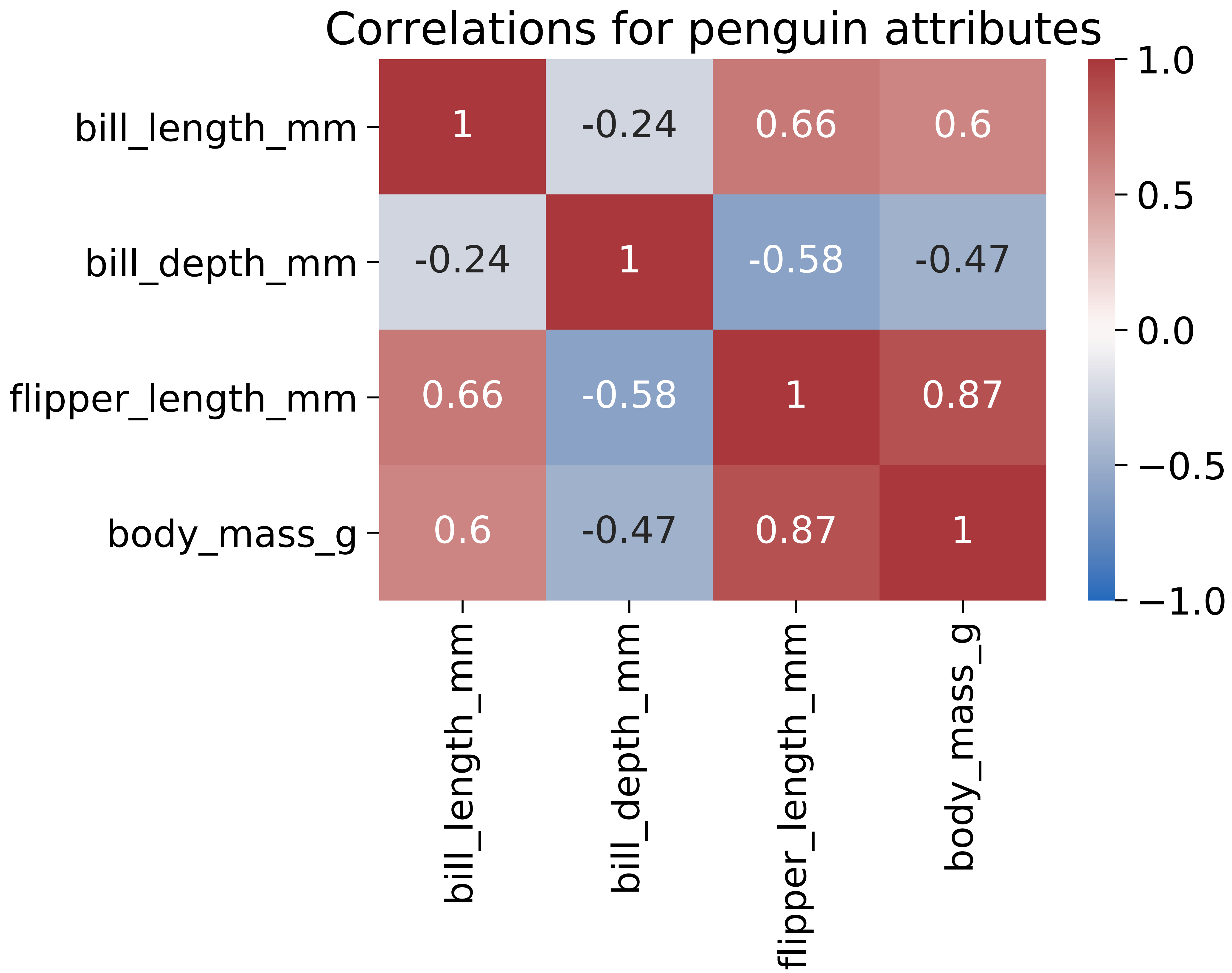
- We can view these statistics split by category
- Any points outside of the *whiskers* are plotted



Plot can be horizontal or vertical

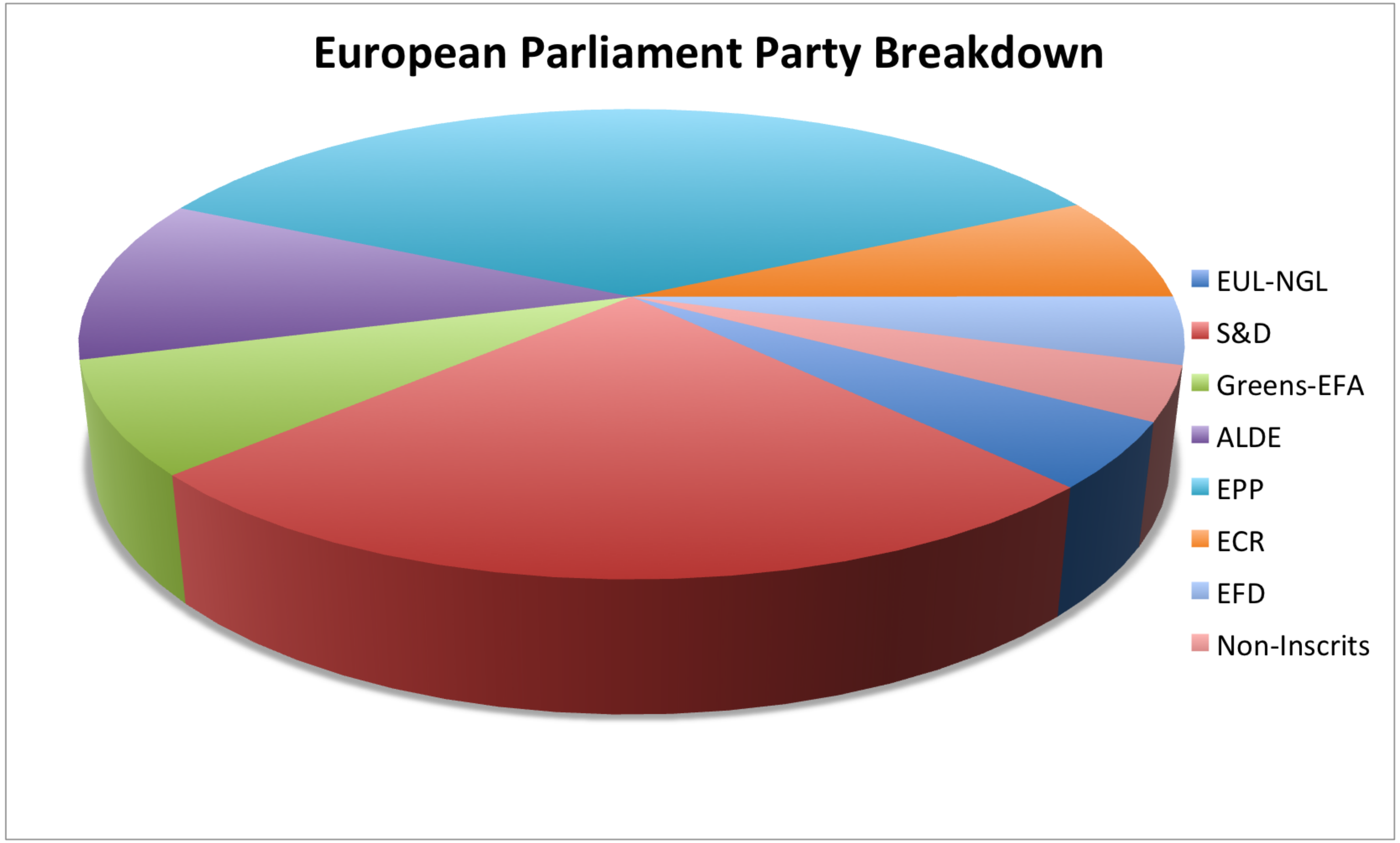
Heat maps

- A matrix with colours to represent intensities of some quantity
- Here we have correlation coefficients of different attributes of penguins



And of course … pie charts

Avoid!



Source: https://www.businessinsider.com/pie-charts-are-the-worst-2013-6?r=US&IR=T

Summary

- We have revised some statistics and seen how they can summarise data
- We have considered Pearson correlations for different pairs of variables
- We have seen examples of good and bad visualisations of data
- We have considered different ways of plotting data