COMP9033
DATA ANALYTICS

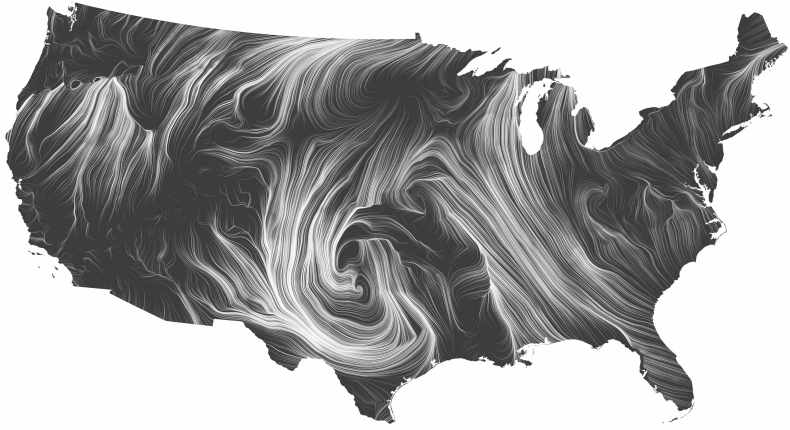
3/12

DATA VISUALISATION

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2017.02.15



Credit: Fernanda Viegas and Martin Wattenberg

Overview

1. Data types:
    - Exploratory data analysis.
    - Types of data.
2. Visual analysis:
    - Scatter/line plots.
    - Histograms.
    - Bar charts.
    - Pie charts.
3. Summary statistics:
    - Central tendency.
    - Dispersion.
4. Detecting anomalies:
    - Outliers.
    - Robust statistics.
    - Graphical techniques.
    - Quantitative techniques.

1. Discovering relationships:
    - Dependence and correlation.
    - Correlation and causation.
    - Quantitative measures.
    - Graphical techniques.
2. Visual cues:
    - What they are, why they're important.
    - Chart types.
    - Effects of misusing visual cues.
3. The data visualisation process:
    - The Four Pillars of Effective Visualisation.
    - Chart formatting.
    - Real world examples.

Discovering relationships

- The existence of a statistical relationship between different data samples is known as *dependence*.
- Two samples are said to be *dependent* when the value of the first *depends* on the value of the second or vice-versa, *e.g.*
  - The number of people wearing shorts on a given day *depends* on the weather.
  - A team's position in a league *depends* on the number of goals they've scored.
  - A student's score on a test (ideally) *depends* on the depth of their knowledge.
- It's possible that multiple dependencies exist: in reality, a student's score on a test depends on more than just knowledge, *e.g.*
  - How well they slept the night before.
  - How much coffee they've drank.
  - The ambient temperature in the test hall.
  - ...lots of other factors!

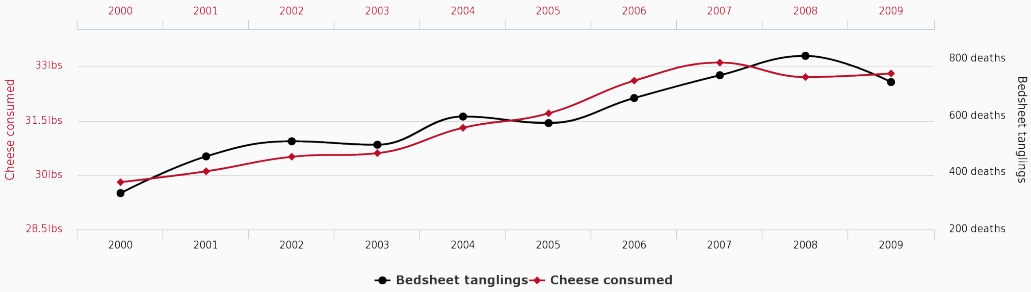
- *Correlation* is a measure of the dependence between two data samples.
- If two samples are *positively* correlated, then their values tend to increase or decrease together:
  - More of *A* leads to more of *B*: Sunshine → ice cream purchases.
  - Less of *A* leads to less of *B*: Eating fewer calories → losing weight.
- If two samples are *negatively* correlated, then the values in one tends to increase/decrease when the values in the other decrease/increase:
  - More of *A* leads to less of *B*: Smoking → lower life expectancy.
  - Less of *A* leads to more of *B*: Less public transport → more congestion.
- If two samples are *uncorrelated*, then there is no co-dependence:
  - More of *A* leads to no change in *B*: Higher taxes → temperature in June.
  - Less of *A* leads to no change in *B*: Population of France → days in a year.

- Dependence does not imply a causal relationship!
  - Buying ice creams does not make the weather sunnier.
  - Living a short life does not make you a smoker.
  - Wearing shorts does not make the weather better.
- If A and B are correlated, then there are six possible explanations:
  1. A is caused by B.
  2. B is caused by A.
  3. A causes B and B causes A.
  4. A and B are both caused by a hidden external factor, C.
  5. A causes C which, in turn, causes B.
  6. A and B are not correlated but, by random chance, they *appear* correlated.
- In most cases, it's not possible to say which of these explanations is true without deeper investigation.

Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

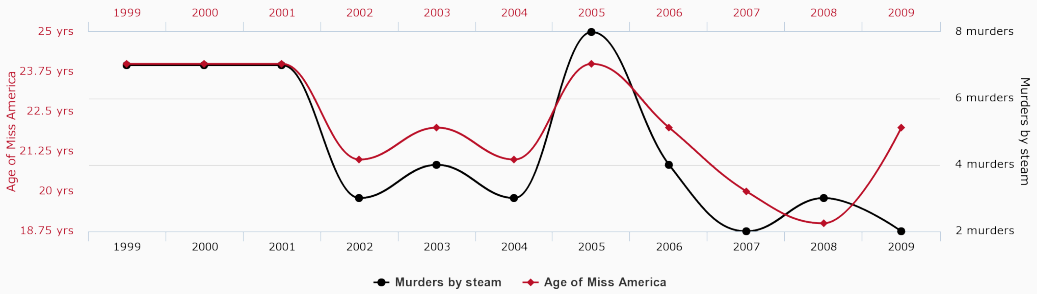


Credit: Tyler Vigen

Age of Miss America

correlates with

Murders by steam, hot vapours and hot objects



Credit: Tyler Vigen

- The *Pearson correlation coefficient* is a commonly used statistical measure of correlation.
- The Pearson correlation coefficient between two data samples, X and Y, is denoted by r_{xy} and is defined as

$$r_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right) \tag{3.1}$$

$$= \frac{1}{n-1} \sum_{i=1}^{n} z(x_i)\, z(y_i). \tag{3.2}$$

- Other kinds of correlation coefficient do exist (*e.g.* Spearman, Kendall), so be careful not to confuse definitions!
- However, in general, the term *correlation coefficient* can be taken to mean Pearson correlation coefficient.

- The Pearson correlation coefficient has a value between -1 and +1, inclusive, *i.e.* $r_{xy} \in [-1, 1]$.
- Positive correlation corresponds to a *positive* value of r_{xy}, *i.e.* $r_{xy} > 0$.
  - If $r_{xy} = 1$, then X and Y have the strongest possible level of positive correlation.
- Negative correlation corresponds to a *negative* value of r_{xy}, *i.e.* $r_{xy} < 0$.
  - If $r_{xy} = -1$, then X and Y have the strongest possible level of negative correlation.
- If $r_{xy} = 0$, then X and Y are uncorrelated.
- One handy property of the coefficient is that $r_{xy} = r_{yx}$, *i.e.* the correlation between X and Y is the same as the correlation between Y and X — the order does not matter.

Q. What is the correlation of the samples $A = \{1, 2, 3\}$ and $B = \{4, 5, 6\}$?

A. First, let's compute the means and standard deviations of the samples using Equation 2.2 and Equation 2.4:

$$\bar{a} = 2,\ \sigma_A = 1. \qquad \bar{b} = 5,\ \sigma_B = 1.$$

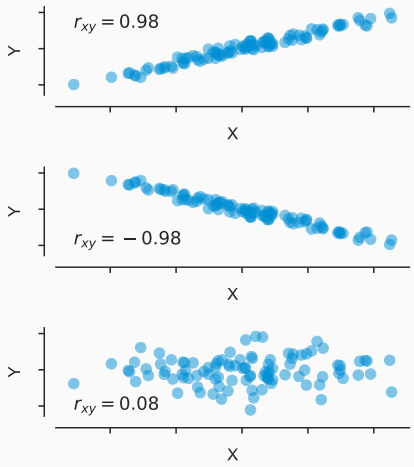
Next, we compute the standard scores of the data points in each sample using Equation 2.8:

$$z(a_1) = -1,\ z(a_2) = 0,\ z(a_3) = 1. \qquad z(b_1) = -1,\ z(b_2) = 0,\ z(b_3) = 1.$$

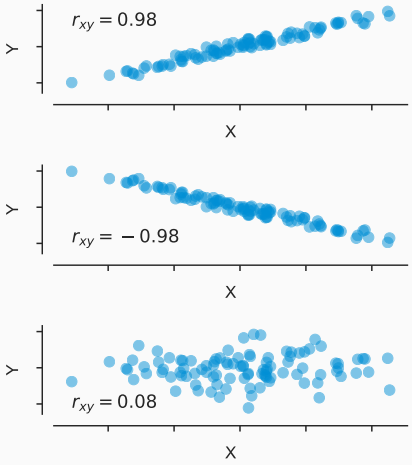
Finally, we compute the correlation coefficient using Equation 3.2:

$$r_{AB} = \frac{1}{n-1} \sum_{i=1}^{n} z(a_i)\, z(b_i) = \frac{((-1) \times (-1)) + (0 \times 0) + (1 \times 1)}{2} = 1.$$

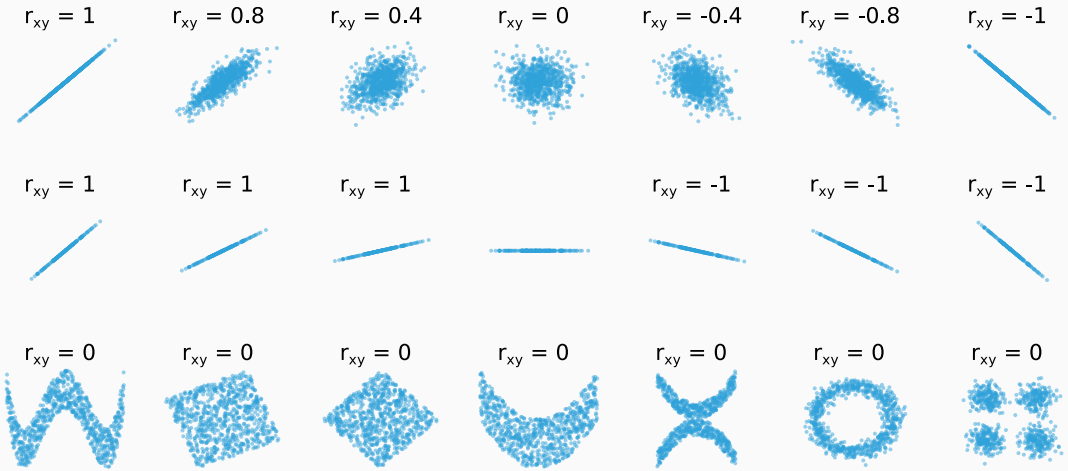
- The Pearson correlation coefficient is a *quantitative* technique for understanding correlation.
- However, we can also use a graphical technique — the scatter plot:
  - Positively correlated data trends linearly upwards.
  - Negatively correlated data trends linearly downwards.
  - Uncorrelated data does not have a linear relationship.



- Pearson correlation measures the strength of *linear* relationships between variables, *e.g.* $y = mx + c$.
- However, other relationships can exist, *e.g.* $y = \sin(x)$, $y = x^2$.
- In many such cases, Pearson correlation will *fail* to show any relationship — however, by combining graphical methods and quantitative methods, we can avoid missing relationships.

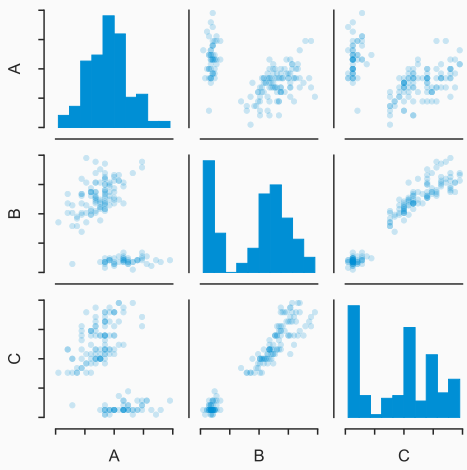


1.11 / EXAMPLE: CORRELATIONS OF VARIOUS RELATIONSHIPS

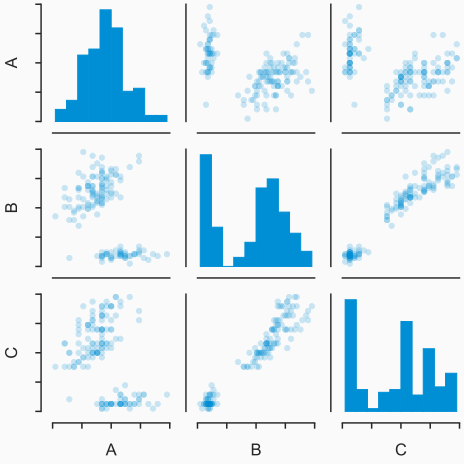


Credit: Denis Boigelot / Wikipedia

- The *scatter plot matrix* is a useful tool for understanding relationships between multiple variables:
  - It consists of a square grid of plots with as many rows/columns as there are variables.
  - The diagonal elements of the grid consist of histograms of single variables (they can also be blank or contain another chart type).
  - The other grid positions are scatter plots of the corresponding row and column variables.



- For instance, the scatter plot matrix to the right shows the relationship between the variables A, B and C:
  - If we want to see how A relates to B, we can look at the second plot in the first row or the first plot in the second row.
  - If we want to see how A relates to C, we can look at the third plot in the first row or the first plot in the third row.
- In this case, it's clear that there is a positive correlation between B and C, but no clear relationship between A and B or A and C.



Visual cues

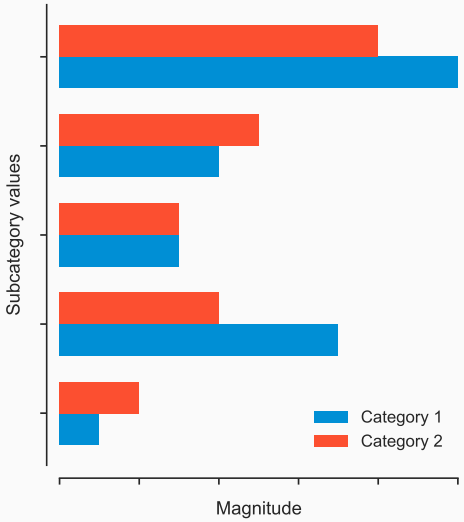
- The primary aim of data visualisation is the *effective* communication of information:
  - Good visualisations make complex data *easier* to understand.
  - Bad visualisations make simple data *harder* to understand.
- One way to make better visualisations is to choose chart types that encode[1] the meaning of our data using appropriate *visual cues*:
  - A visual cue graphically encodes data with shapes, colours, sizes, *etc*.
  - Generally, visual cues are self-explanatory — we intuitively understand what they mean, *e.g*. the length of a bar in a bar chart conveys magnitude.
  - If we choose visual cues well, we can minimise clutter and maximise intuitive understanding of our visualisations.

[1] For more information, see bit.ly/2kUpGQA.

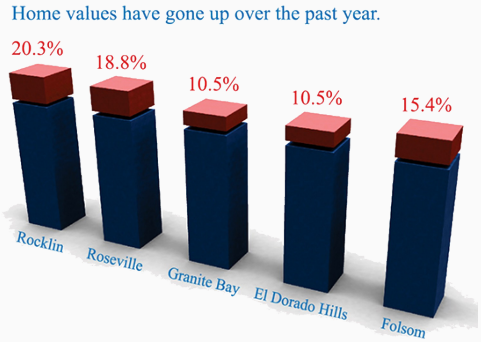
2.2 / EXAMPLES OF VISUAL CUES

VISUAL CUE	NUMBER OF VARIABLES	EXAMPLE USAGE
length	large	the length of bars in a bar chart
size/area	large	the size of bubbles in a bubble chart
position/placement	large	the placement of bubbles in a bubble chart
connection	large	edges between nodes in a network graph
angle	moderate	the angle of slices in a pie chart
shape/icon	moderate	highlighting points in a scatter plot
colour/saturation	small	the colour of bars in a bar chart
line pattern	small	highlighting different lines in a line plot
line weight	small	highlighting different lines in a line plot
line endings	small	highlighting the direction of trends in a line plot

- Bar charts encode meaning of categoric data using *length* and *colour*:
  - The length of a bar indicates the magnitude of the corresponding variable.
  - The colour of a bar indicates the category of the corresponding variable, *e.g.* when plotting multiple series.

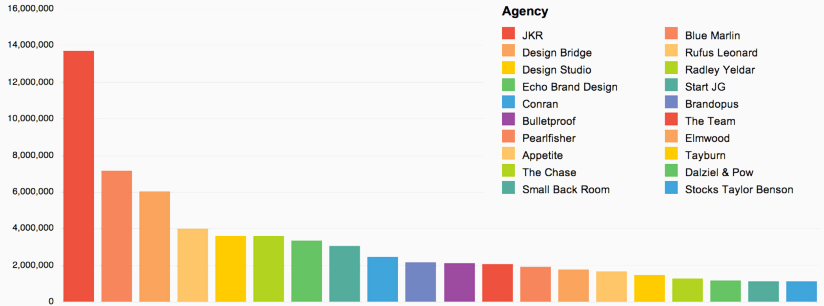


- If we misuse a visual cue, we can make our graphics harder to understand.
- For instance, in this case, the value that the left-most bar represents is almost twice the value of the middle bar.
- However, because each bar has been placed on a “pedestal”, this difference is not apparent at first glance.



Credit: Infographic Marketing

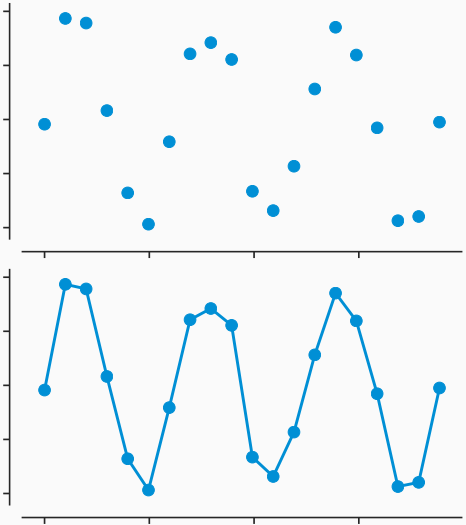
2.5 / EXAMPLE: MISUSING COLOUR



Credit: DesignStudio

- In this case, colour has been misused — some bars have very similar colours (*e.g.* the second and third from the left), while the colours themselves repeat halfway across making the true meaning of the chart unintelligible.

- Scatter/line plots encode meaning through *position*, *colour*, *shape/icon*, *line pattern*, *line weight* and *line ending*:
  - The position of points indicates their value.
  - The colour of points/line indicates their meaning, *e.g.* when plotting multiple series.
  - Varying the colour/styles of points (*e.g.* circles, squares, crosses) can emphasise different trends or subgroups.
  - Varying line properties can help to differentiate when plotting multiple series.
  - Adding a line ending (*e.g.* an arrow) shows directionality/order.

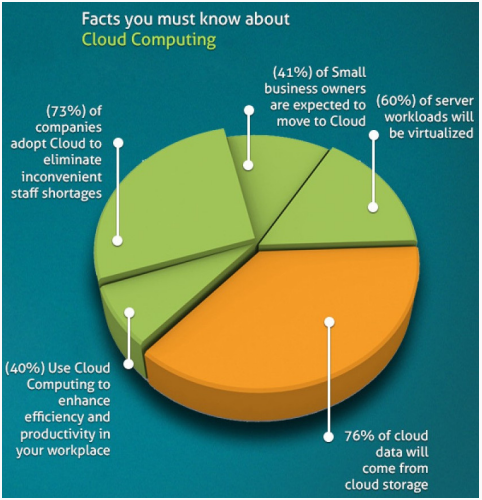


- The positions of scatter points (and lines) make the trend of the data immediately obvious.
- However, if the distances between points is warped, or if points are excluded from the trend line, the meaning of the data becomes distorted.
- In the chart to the right, the *x* axis distances look even, but represent different time gaps.
- Also, as the data is quarterly, there should be sixteen points in total (four points per year), not one from a different month over four years.

Credit: Fox News

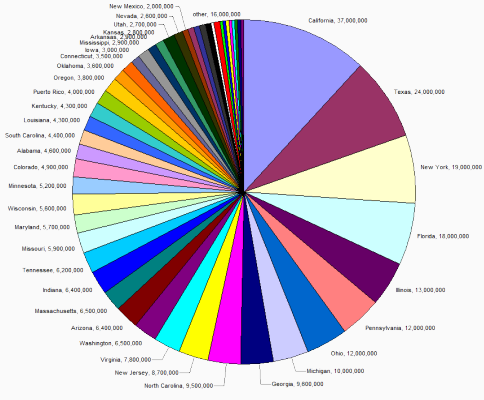
- Pie charts encode meaning using *angle* and *colour*:
  - The angle of a pie slice indicates the magnitude of the corresponding variable.
  - The colour of a pie slice indicates the category of the corresponding variable.

- Pie charts only make sense when the magnitudes of the slice variables sum logically.
- After all, a pie chart is a circle, so there are just 360° to share.
- When slice values don't add up, neither does the visualisation.



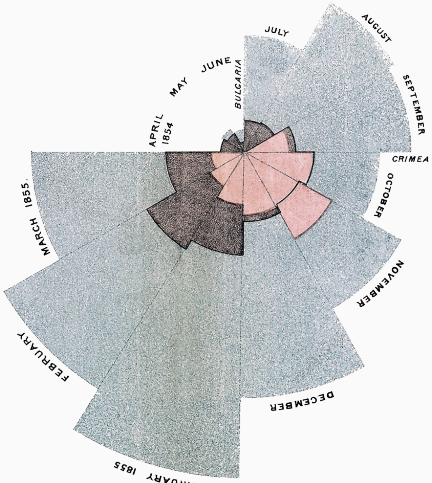
Credit: Motion Wave

- As you add more variables to the chart, the angles of each slice becomes smaller.
- Eventually, you will reach a point where it's no longer easy to understand the magnitude represented by a slice.
- The problem is compounded by the fact that you can't choose a large number of easily distinguishable colours.



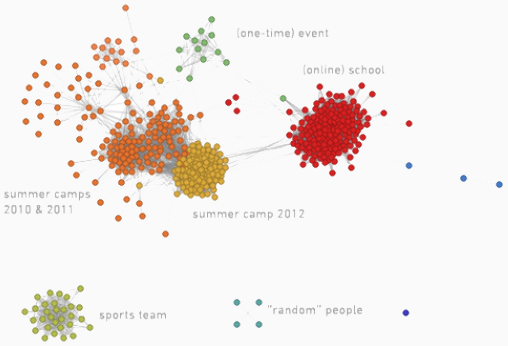
Credit: Mikael Häggström/Wikipedia

- A polar area plot is a variation on the pie chart that relies on *colour* and *length*:
  - The colour of a sector indicates the category of the corresponding variable.
  - The length of a sector indicates the magnitude of the corresponding variable.
- Polar area charts have no dependency on angle — the angle of each sector is equal.



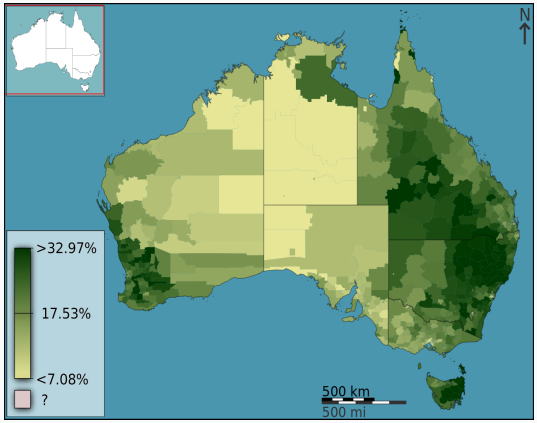
Credit: Florence Nightingale / Wikipedia

- Network graphs encode the meaning of relational data using *connection*, *colour* and *area*:
  - The connections between nodes communicate relationships.
  - The colour of nodes indicates the category to which they belong.
  - The area of the nodes (not used to the right) indicates the magnitude of the value of the corresponding variable.



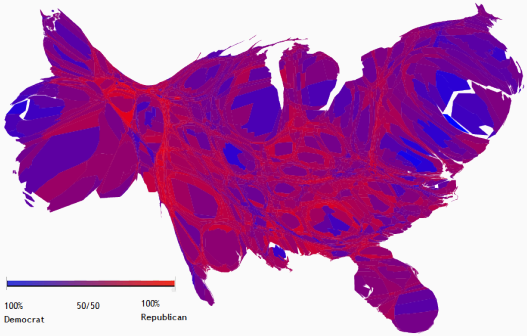
Credit: Stephen Wolfram

- The chloropleth map encodes the meaning of spatial data through *colour*:
  - The colour of a region of the map indicates the magnitude of the variable being measured.
- Other kinds of map also encode meaning using colour, *e.g.* category membership.

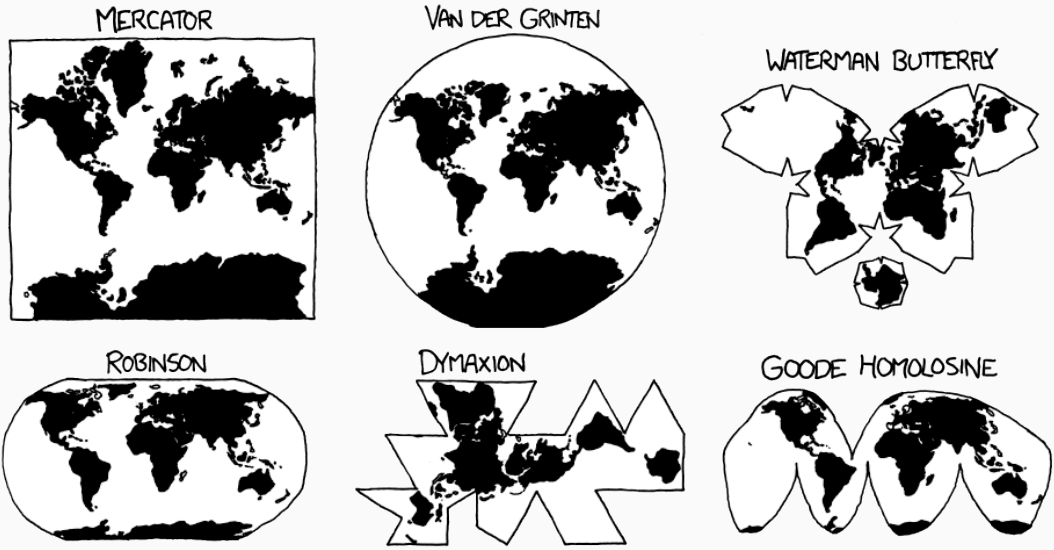


Credit: Toby Hudson / Wikipedia

- The cartogram encodes the meaning of spatial data using *colour* and *area*:
  - Like the chloropleth map, the colour of a region indicates the value of the variable being measured.
  - The area of regions is usually scaled, producing a warped effect, to indicate the value of a second variable, typically population density.



Credit: M. E. J. Newman / Wikipedia

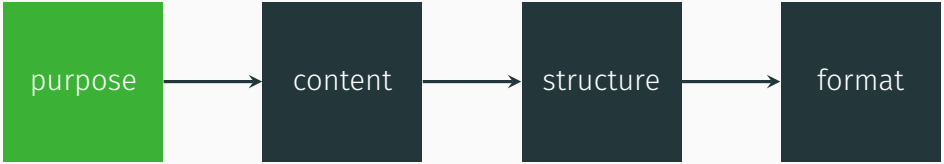


Credit: Randall Munroe/xkcd

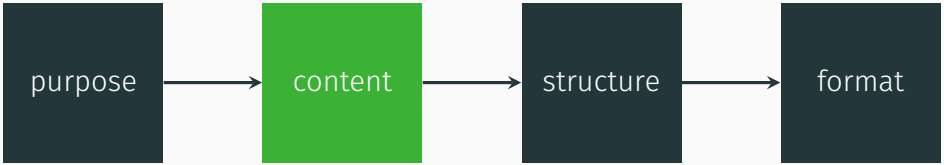
The data visualisation process

- Like processes for data analysis, there are also processes for data visualisation.
- One such process, known as the *Four Pillars of Effective Visualisation*[2], emphasises the following sequence:
  1. Purpose: *why* are you creating your visualisation?
  2. Content: *what* are you going to visualise?
  3. Structure: *how* are you going to visualise it?
  4. Formatting: *who* is your audience?

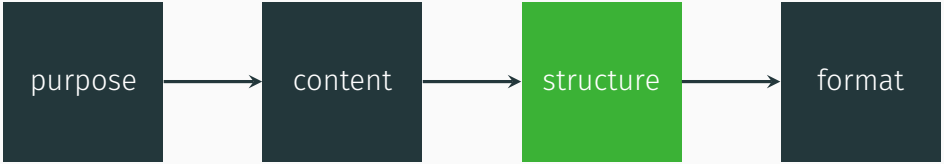
[2] For more information, see bit.ly/2llzal5.



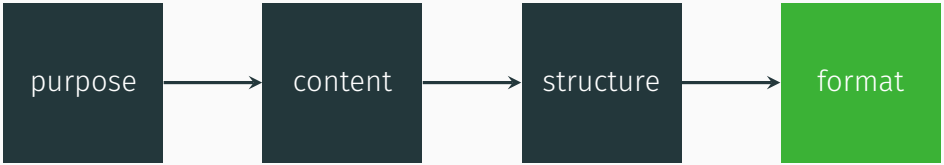
- The first stage in designing an effective visualisation is to define its purpose, *i.e.*
  - What is your purpose?
  - What is the aim of your visualisation?
  - What information are you trying to convey?



- The second stage in designing an effective visualisation is to decide what its content will be, *i.e.*
  - What data are you going to visualise?
  - Do you have enough data?
  - Do you have too much data? Do you need to visualise it all or just a subset?
  - Are there relationships in the data that support your purpose?



- The third stage in designing an effective visualisation is to decide on a method to visualise your data with; this generally depends on:
  1. The data type you are working with.
  2. The properties of your data, which are encoded by *visual cues*.
- Choosing a poor structure makes your visualisation more difficult to understand, so this is a crucial step.

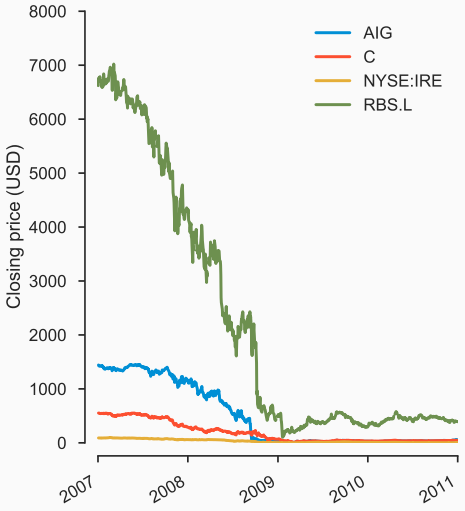


- The final stage in designing an effective visualisation is to decide how much additional formatting is required.
- This is also a crucial step, as it determines the amount of time you should spend touching up your visualisation once the first three steps are complete.
- Generally, this depends on the intended audience of the visualisation, *i.e.* *who* will be viewing your image?

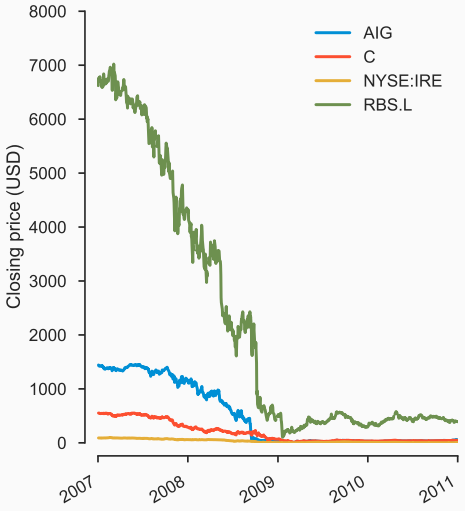
- Why consider your audience? Because technical and non-technical audiences have different requirements!
- If your audience is technical (*e.g.* fellow team members, subject matter experts), then a "quick and dirty" visualisation might be best:
  - Misunderstandings can usually be cleared up quickly.
  - If speed is a priority, then image quality is usually not.
- If your audience is non-technical (*e.g.* management or customers), then you might want to spend more time making things look good:
  - Well designed graphics are easier to interpret and understand, and so can save time, questions and frustration.
  - If your visualisation looks good, you look good!

- So, what should you consider when formatting your visualisation?
  - Whether to label graph axes and, if so, with what.
  - Whether to include a plot legend and, if so, of what kind.
  - What colour scheme to use, if colour is used (*e.g.* should it be colour blind friendly?).
  - Whether to include additional annotation to highlight important features (*e.g.* the month with the highest sales).
  - Whether to use grid lines, which can make visual comparison easier, but also add unwanted clutter.
  - What aspect ratio to use.
  - What font to use.
  - ...essentially, any form of visual polish that makes your graphic easier to interpret!
- Be careful not to visually clutter your graphic while formatting it — this will undo all the benefits!

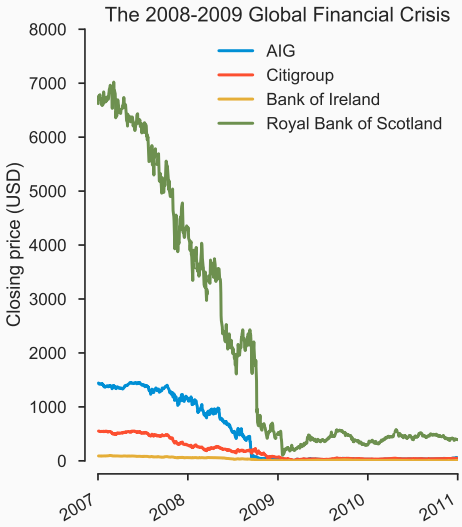
- The image to the right shows the closing stock prices of four major companies around the time of the 2008-2009 global financial crisis.



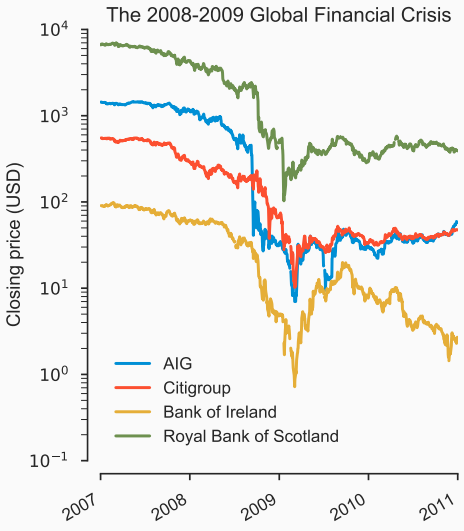
- Some formatting has already been applied:
  - The trend lines have been given distinct colours, to make them easily distinguishable.
  - The *y* axis has been titled to make its meaning clearer.
  - The *x* axis has not been titled; its tick labels have been formatted as dates (and rotated to fit) instead.
  - The plot has been given a legend, so that we can look up what each trend line represents.



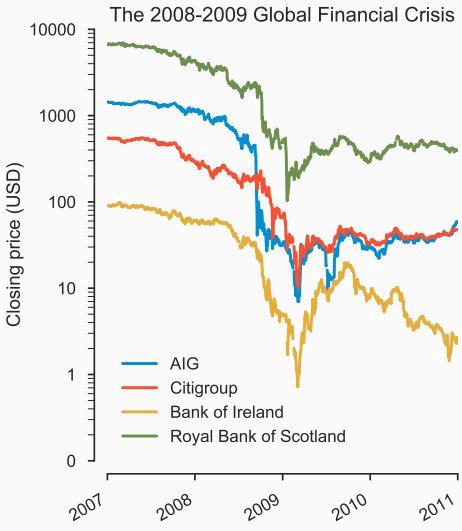
- We can take this further though!
  - Adding a title makes the *purpose* of our chart immediately clear.
  - We can also remove the stock tickers and replace them with the names of the companies they represent to make the legend easier to understand.



- Converting the *y* axis scale from linear to logarithmic emphasises the exponential changes in the magnitude of the data:
  - Small changes are more apparent.
  - Big changes are still clear.
  - It's now clearer that all of the trends have experienced a similar phenomenon (*i.e.* emphasise the *purpose* of the image).
- In this case, the legend is also repositioned so as not to overlap with any of the trend lines.



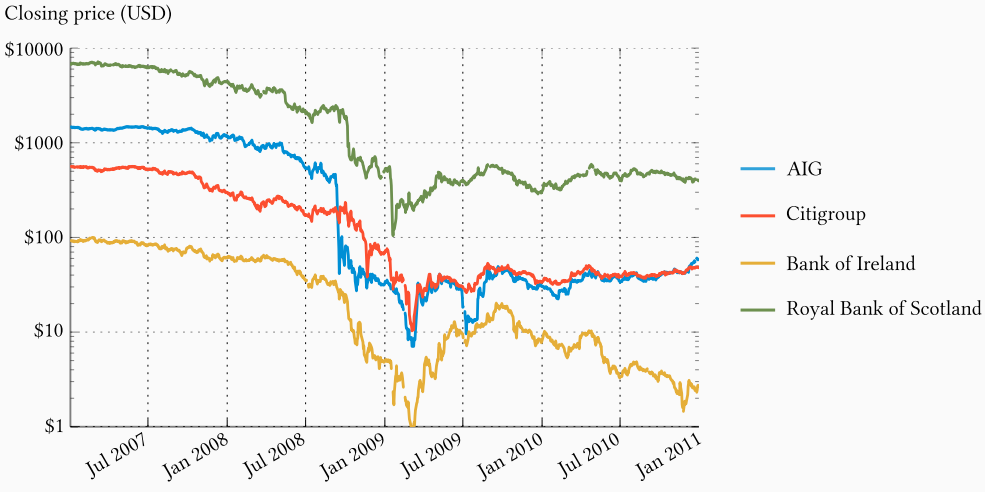
- Simplifying the *y* axis labels adds further clarity:
  - Scientific notation for numbers (*e.g.* 10^4) is the norm in some sectors, organisations and businesses, but not in others.
  - If our audience is technical (*e.g.* engineers, statisticians), then it may be fine to use scientific notation — the extra detail may be appreciated.
  - If our audience is not scientific, then using natural numbers may make the *purpose* of the graphic more readily understood.



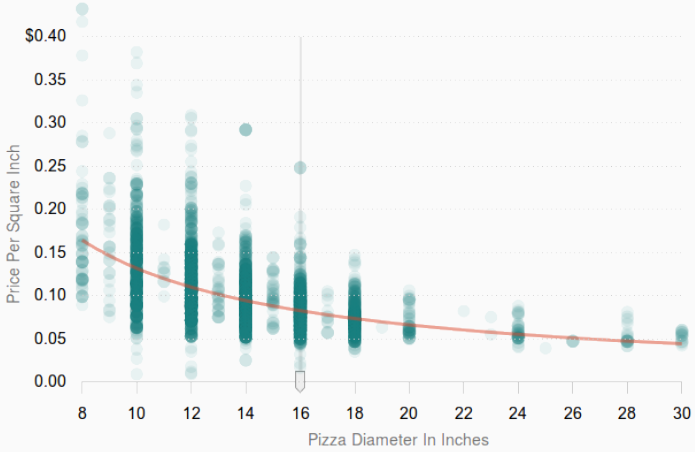
- Often, you can do almost all the formatting you'll need using code (*e.g.* matplotlib, pandas, seaborn).
- However, while languages like Python and R have a good selection of graphic manipulation libraries, sometimes we want extra *oomph*.
- In such cases, the convention is to export your image in vector graphic form (*e.g.* PDF, SVG) and edit it directly using an image manipulation tool, such as Adobe Illustrator or Inkscape.
- This gives you much finer grain control over layout, colour and fonts, and makes it significantly easier to produce production quality images.
- However, the additional effort required is usually costly (in terms of time) and so is not appropriate for every situation - know your audience and act accordingly!

The 2008-2009 Global Financial Crisis

How the mighty have fallen: a selection of international financial institutions and how they were affected.



3.15 / EXAMPLE: PIZZA PRICES



Credit: NPR/Planet Money

Purpose To show how pizza prices vary *and* how they vary with size.

Content 74,476 prices from 3,678 different pizza places.

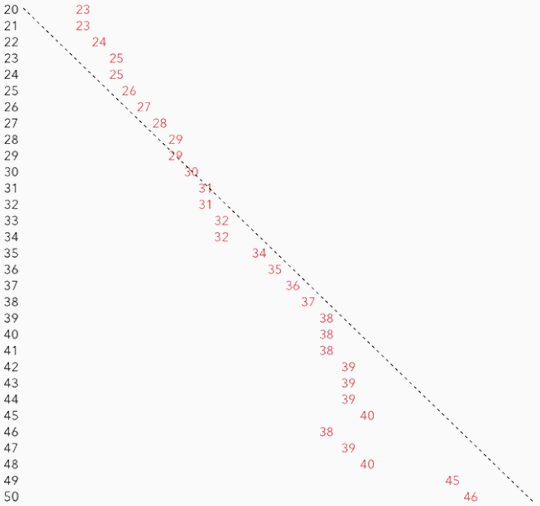
Structure A scatter plot of prices versus size.

Format The best fit line (in red) shows the central tendency of the prices, while the scatter points (torquoise) show the dispersion. Clusters of similar price points are emphasised by setting the opacity of individual points to be low.

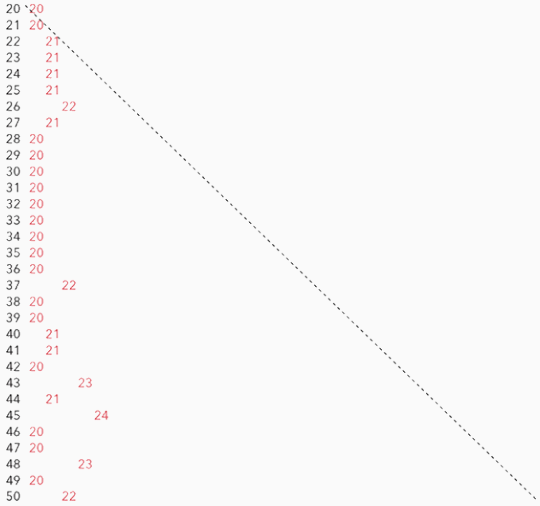
The axes titles, labels and ticks are minimal, but still help to clarify the meaning of the image. In particular, only the upper most price on the *y* axis is given in dollars, which communicates the currency of the remaining prices, but minimises visual clutter.

3.17 / EXAMPLE: OK CUPID

a woman's age vs. the age of the men who look best to her



a man's age vs. the age of the women who look best to him



Credit: Christian Rudder / Five Thirty Eight

Purpose To highight differences in the behaviour of OK Cupid users.

Content The ages and genders of OK Cupid users and the ages and genders of the users whose profiles they viewed.

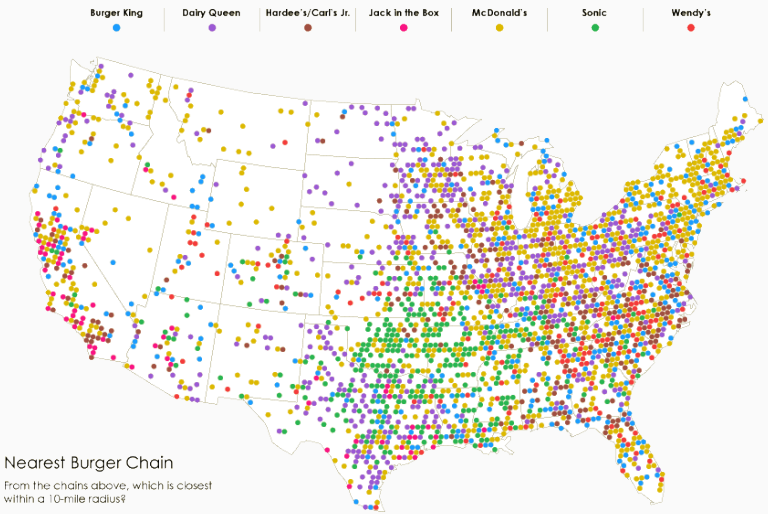
Structure Horizontal bar plots of the average age of viewed profiles (*x* axis) per age group category (*y* axis).

Format The plot title also acts as a legend, and so the meaning of red and black are immediately clear.

No visible bars are used: instead numerical markers denote where the tops of the bars would be. This negates the need for a labelled *x* axis and also means that the exact numerical value for each bar is immediately clear.

The dotted guide line illustrates idealised equality, allowing us to spot deviations quickly and quantify their magnitude easily.

3.19 / EXAMPLE: BURGER PLACE GEOGRAPHY



Credit: FlowingData

Purpose To illustrate the density and variety of fast food chains in the United States.

Content The locations of all seven fast food chain outlets.

Structure A scatter plot of the nearest fast food outlet within a given radius (this would have to be computed from the original data somehow), backed by a map of the US.

Format The choice of distinctive colours means that neighbouring points representing different burger chains are easily differentiated[3].
The plot title and legend are not in their conventional positions, *i.e.* above and to the right/below. However, the legend is easy to read horizontally and aligns neatly with the flat top of the map. The placement of the title is similarly tidy, and adds to the overall minimal feel of the graphic.

[3] For more information on choosing colours for maps, see ColorBrewer.

Summary

- Lots of material this week!
  - Dependence, correlation and causation.
  - Discovering relationships — graphical and quantitative techniques.
  - Visual cues and how to use them.
  - A process for data visualisation — why, what, how and who?
- This week's lab focuses on how to apply visual EDA techniques with pandas:
  - Bar charts.
  - Pie charts.
  - Histograms.
  - Scatter plot matrices.
- Next week: cleaning and transforming data!

1. Khan Academy. *Data and statistics.* (bit.ly/1DZTQpA)
2. R Psychologist. *Interpreting correlations: an interactive visualization.* (bit.ly/1FAqVXb)
3. Yau, Nathan. *Data points: visualization that means something.* John Wiley & Sons, 2013. (bit.ly/2k8TqWR)
4. Tufte, Edward. *The Visual Display of Quantitative Information.* Graphics Press, 2001. (bit.ly/2kAU2Ic)