COMP9033
DATA ANALYTICS

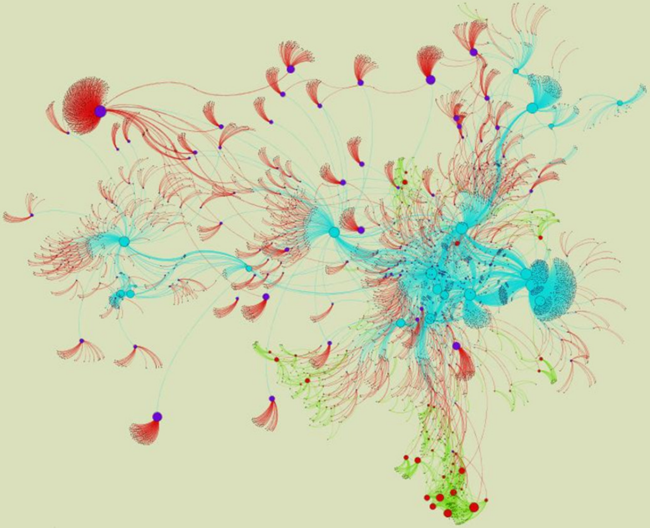
2/12

EXPLORATORY DATA ANALYSIS I

DR. DONAGH HORGAN

DEPARTMENT OF COMPUTER SCIENCE
CORK INSTITUTE OF TECHNOLOGY

2018.02.06



Credit: Elijah Meeks / Stanford University

Overview

1. Introduction to data analysis:
    - What is it?
    - How does it work?
    - Real world examples.
2. Module outline:
    - Overview of topics.
    - Marking scheme.
    - Lab work.
    - Project work.
    - Contact information.
3. Data analysis processes:
    - What are they?
    - Why use them?
    - How do they work?
    - Which one to use?
4. Data sampling:
    - What is it?
    - Why is it important?
    - How to do it?

0.2 / THIS WEEK

1. Exploratory data analysis:
    - What is exploratory data analysis?
    - Why is it important?
    - Data types.
2. Visual data analysis:
    - Chart types.
    - Visual cues — what they are, why they're important.
    - The Four Pillars of Effective Visualisation.
3. Distributions:
    - Histograms.
    - The normal distribution.
    - Other kinds of distribution.
    - Visualising category distributions.

Exploratory data analysis

- The term *exploratory data analysis* (EDA) refers to a commonly used approach for analysing data sets.
- Broadly, the aims of EDA are to:
    - Become familiar with the data to be analysed.
    - Uncover hidden structure in the data.
    - Determine whether the data contains important variables.
    - Detect anomalies in the data.
    - Test assumptions about the data.
- Generally, EDA is carried out *after* sampling the data but *before* cleaning and/or transforming it (recall SEMMA/CRISP-DM from Lecture 01).

- EDA techniques consist of both *quantitative* and *graphical* methods, *e.g.*
  - Categorising the type of the data, *e.g.* time series, geographic coordinates.
  - Visualising the behaviour of the data, *e.g.* time series plot, histogram.
  - Summarising the behaviour of variables, *e.g.* typical values, ranges.
  - Detecting outliers and anomalies, *e.g.* snow in summertime.
  - Determining whether the data follows a particular distribution, *e.g.* the normal distribution.
  - Discovering or verifying relationships between variables, *e.g.* more sunshine $\rightarrow$ more icecream sales.
  - Finding groups or categories within the data, *e.g.* distinct species in a sample of animals.

- The *type* of data describes its content, *e.g*. whether it is numeric, categoric, a time series, GPS coordinates, *etc*.
- Defining the type of data you are working with is important as it affects the kind of techniques you can use throughout the remainder of the analysis process, *e.g*.
    - We can't compute the average of a set of categories (*e.g*. {Dog, Cat, Dog}).
    - Time series data may need to be treated sequentially in order to preserve certain chronological properties.
    - Spatial data may need to be transformed into a common coordinate reference system (*e.g*. WGS84).
- Data can have more than one type — in such cases, you should determine the type that is most relevant to the analysis you are carrying out.

Quantitative Numeric data with no inherent order or dependencies, *e.g.*

- Currency.
- Temperature.
- Population.

Categoric Data consisting of unordered groups of items, *e.g.*

- Breeds of dog.
- Car manufacturers.
- Countries with the Euro currency.

Ordinal Data with an intrinsic order (*i.e.* the sequence matters), *e.g.*

- Relative popularity of political parties (also numeric).
- Finishing positions in a race (also categoric).
- Countries with the Euro currency in GDP order (also categoric).

Spatial Data measured with respect to location, *e.g.*

- GPS coordinates (also numeric).
- Post codes (also categoric).
- Addresses (text, can be converted to coordinates or post codes).

Temporal Data measured with respect to time, *e.g.*

- Currency fluctuation over time (also numeric).
- World Cup winners (also categoric).
- GPS trace of running route (also spatial).

Relational Data with an inherent structure, *e.g.*

- Social network contacts.
- Organisation chart hierarchy.
- Commonly purchased groups of items.

Visual data analysis

- Visual analysis can give us an intuitive understanding of data quickly:
    - Quantitative techniques (*e.g.* statistics) give us precise numerical answers.
    - However, digesting large amounts of quantitative data can be overwhelming.
    - Graphical techniques allow to us to get a high level "feel" for what's going on.
    - However, graphical techniques do not give the same level of precise numerical detail as quantitative techniques.
    - A combination of both is typically the best approach.

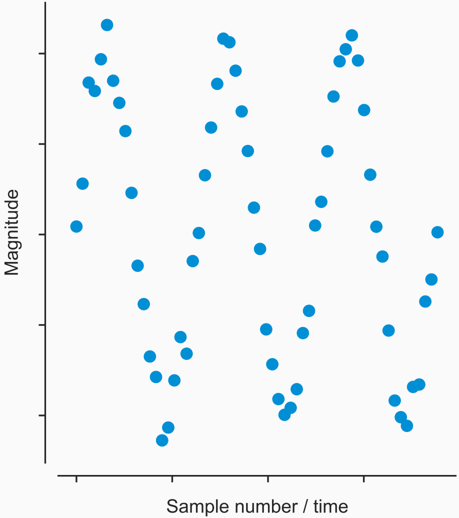
- The primary aim of data visualisation is the *effective* communication of information:
  - Good visualisations make complex data *easier* to understand.
  - Bad visualisations make simple data *harder* to understand.
- One way to make better visualisations is to choose chart types that encode[1] the 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.

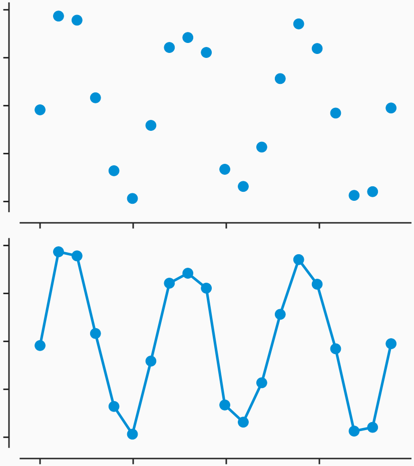
2.3 / EXAMPLES OF VISUAL CUES

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

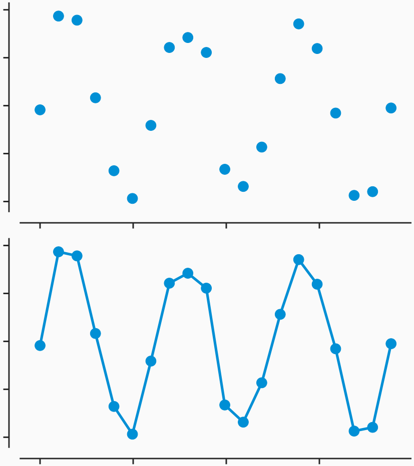
- Scatter plots can help us to understand trends in time series and other ordered quantitative data samples:
  - The *x*-axis measures the *sample order* (*e.g.* time) of the data points in the sample.
  - The *y*-axis measures the *magnitude* of the values in the sample.



- If the number of data points in the sample is small, then it might be difficult to determine a trend visually.
- If the number of data points is not too small, we can use a line plot to help:
    - The axes in a line plot work the same way as in a scatter plot.
    - Consecutive points are connected by a trend line.



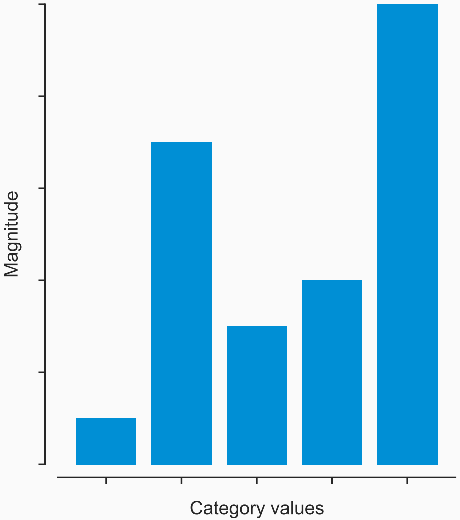
- Meaning can be encoded in a number of ways:
  - The position of points indicates their value.
  - The colour of points/lines indicates their meaning, *e.g.* when plotting multiple series.
  - Varying the colour/styles of points (*e.g.* circles, squares, crosses) can emphasise different trends or subgroups.
  - Varying line properties can help to differentiate when plotting multiple series.
  - Adding a line ending (*e.g.* an arrow) shows directionality/order.



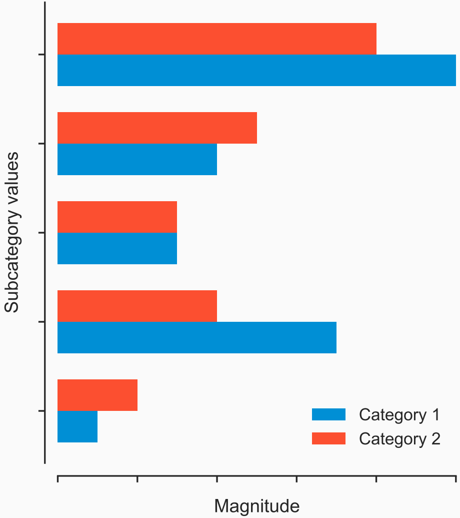
- If the distances between points is warped, or if points are excluded from the trend line, the meaning of the data becomes distorted!
- In the chart to the right, the *x* axis distances look even, but represent different time gaps.
- Also, as the data is quarterly, there should be sixteen points in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

Credit: Fox News

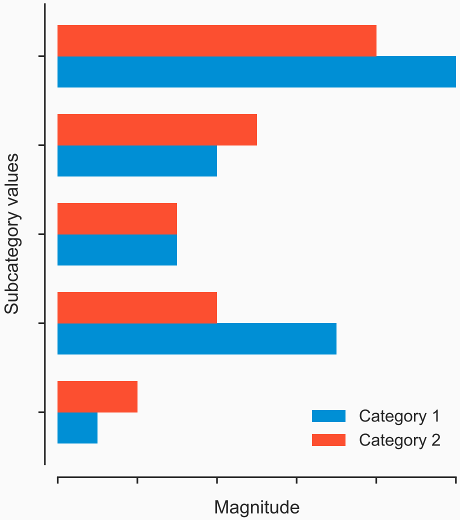
- Bar charts are a useful way to visualise the magnitude and relative proportion of quantities in a categoric sample:
    - The *x*-axis measures the *category value* of the data points in the sample.
    - The *y*-axis measures the *magnitude* of the value of the corresponding category.



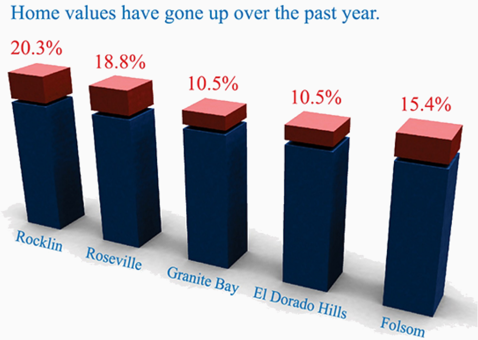
- Bar charts can also be displayed horizontally — we just have to swap the *x* and *y* axes.
- Category hierarchies can be compared by using different colours (*e.g.* number of Olympic medals won by the US and UK in different events).
- It's important to include a legend in this case, so that the meaning of the colours can be distinguished.



- 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.

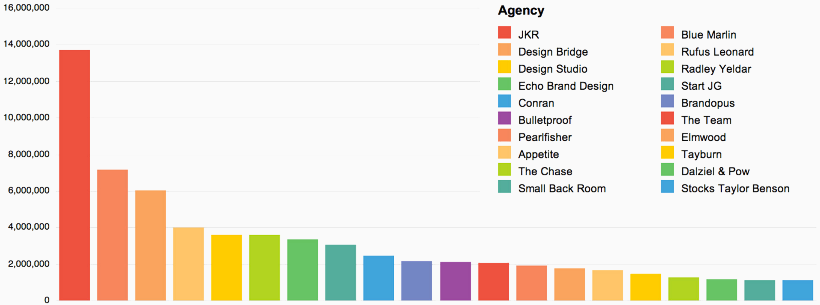


- If we misuse a visual cue, we can make our graphics harder to understand.
- For instance, in this case, the value that the left-most bar represents is almost twice the value of the middle bar.
- However, because each bar has been placed on a “pedestal”, this difference is not apparent at first glance.



Credit: Infographic Marketing

2.12 / EXAMPLE: MISUSING COLOUR



Credit: DesignStudio

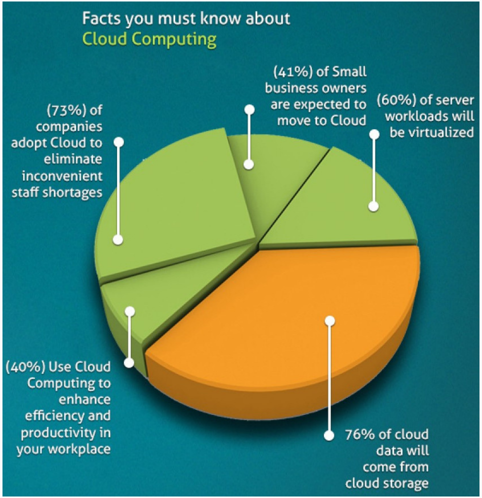
- In this case, colour has been misused — some bars have very similar colours (*e.g.* the second and third from the left), while the colours themselves repeat halfway across making the true meaning of the chart unintelligible.

- Pie charts can also be useful when visualising proportions in a categoric sample:
  - The colours of the sections represent the *category values* of the data points in the sample.
  - The angles of the sections measure the *magnitude* of the value of the corresponding category.

- Pie charts can be useful, but are often difficult to interpret:
    - The difference in the lengths of angles can be harder to discern than the difference in the height of bars.
    - For instance, in the image on the right, which is larger — pink or green?
- For more information, see read.bi/1MIkvcB.

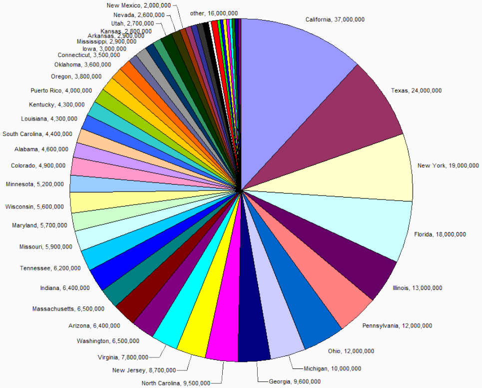
2.15 / EXAMPLE: MISUSING ANGLES

- 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.

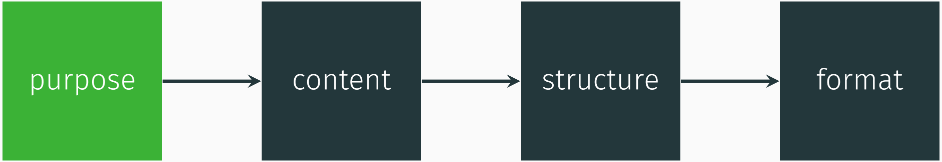


Credit: Mikael Häggström / Wikipedia

- Creating effective visualisations can be tricky at times — there are lots of factors to consider.
- Following a process model can help us to remember all the essential steps and considerations.
- One such process, known as the *Four Pillars of Effective Visualisation*[2], emphasises the following steps:
    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?

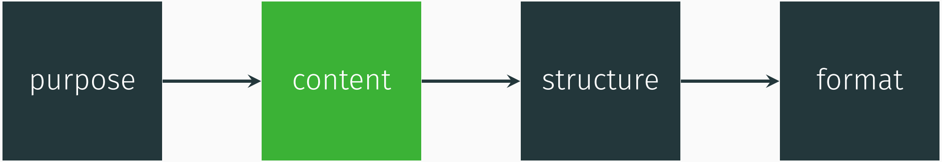
[2] For more information, see bit.ly/2llzal5.

2.18 / THE DATA VISUALISATION PROCESS: PURPOSE



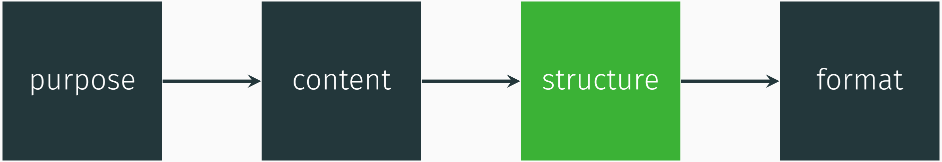
- The first stage in designing an effective visualisation is to define its purpose, *i.e.*
  - What is your purpose?
  - What is the aim of your visualisation?
  - What information are you trying to convey?

2.19 / THE DATA VISUALISATION PROCESS: CONTENT



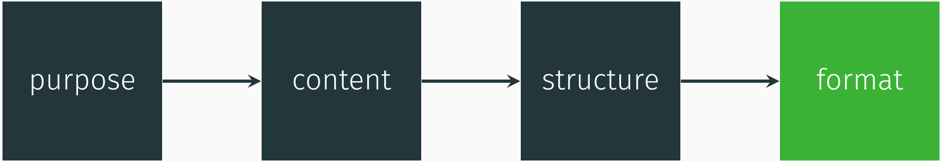
- The second stage in designing an effective visualisation is to decide what its content will be, *i.e.*
  - What data are you going to visualise?
  - Do you have enough data?
  - Do you have too much data? Do you need to visualise all of it or just a subset?
  - Are there relationships in the data that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

2.20 / THE DATA VISUALISATION PROCESS: STRUCTURE



- 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, which should decide what chart type to use.
    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.

2.21 / THE DATA VISUALISATION PROCESS: FORMAT

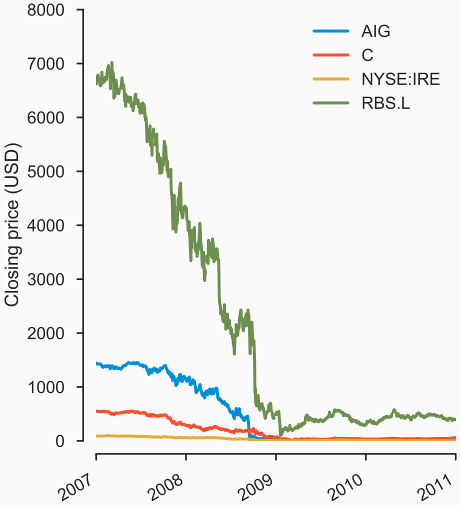


- The final stage in designing an effective visualisation is to decide how much additional formatting is required.
- This is also a crucial step, as it determines the amount of time you should spend touching up your visualisation once the first three steps are complete.
- Generally, this depends on the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

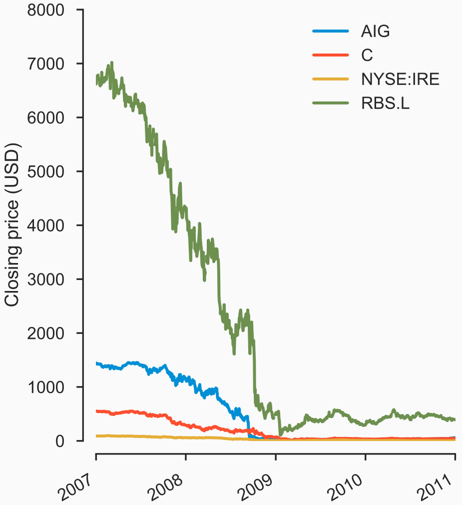
- Why consider your audience? Because technical and non-technical audiences have different requirements!
- If your audience is technical (*e.g.* fellow team members, subject matter experts), then a "quick and dirty" visualisation might be best:
    - Misunderstandings can usually be cleared up quickly.
    - 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!

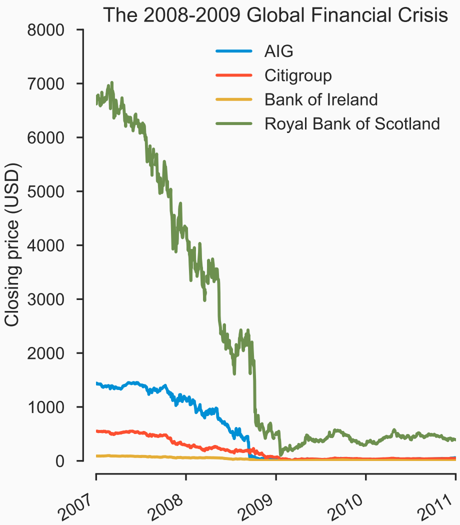
- The image to the right shows the closing stock prices of four major companies around the time of the 2008-2009 global financial crisis.



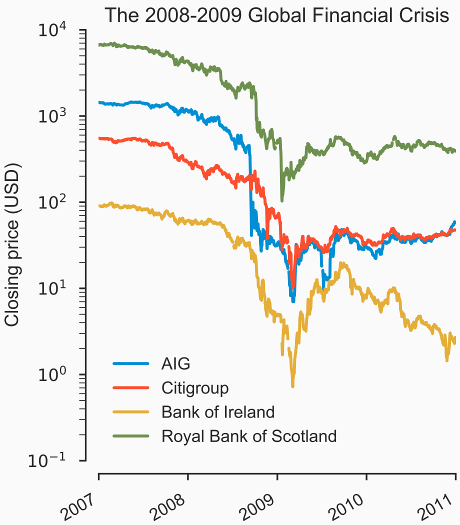
- Some formatting has already been applied:
    - The trend lines have been given distinct colours, to make them easily distinguishable.
    - The *y* axis has been titled to make its meaning clearer.
    - The *x* axis has not been titled; its tick labels have been formatted as dates (and rotated to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



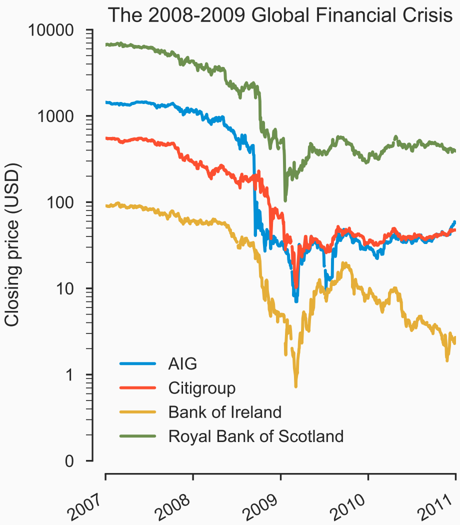
- We can take this further though!
  - Adding a title makes the *purpose* of our chart immediately clear.
  - We can also remove the stock tickers and replace them with the names of the companies they represent to make the legend easier to understand.



- 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).
- The legend is also repositioned so as not to overlap with any of the trend lines.



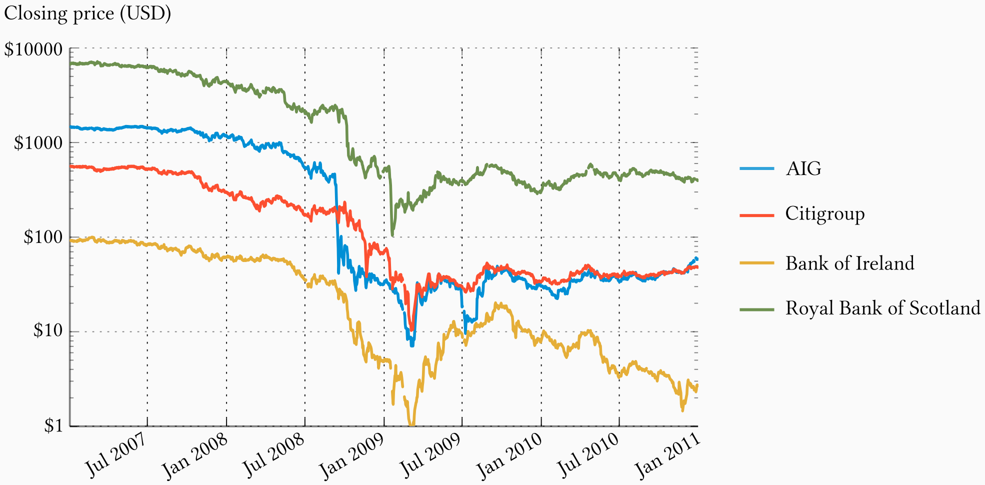
- Simplifying the *y* axis labels adds further clarity:
    - Scientific notation for numbers (*e.g.* 10^4) is the norm in some sectors, organisations and businesses, but not in others.
    - If our audience is technical (*e.g.* engineers, statisticians), then it may be fine to use scientific notation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

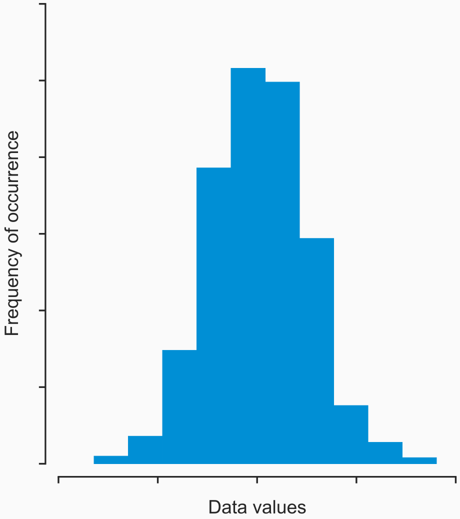
How the mighty have fallen: a selection of international financial institutions and how they were affected.



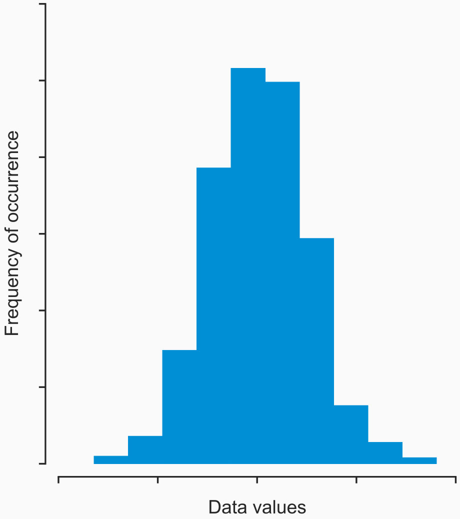
Distributions

- A *distribution* is a measurement of the frequency or likelihood of occurrence of a given value, *e.g.*
    - In an Amazon product review, how many users gave one star, two stars, *etc.*?
    - What are the chances of winning a lottery with a given combination of numbers?
    - What is the temperature range in Madrid in September? What temperature is most likely?
- We can compute distributions by counting how often different values occur in our data sample.
- Examining the distribution of the variables in your data is an essential step in any analysis!

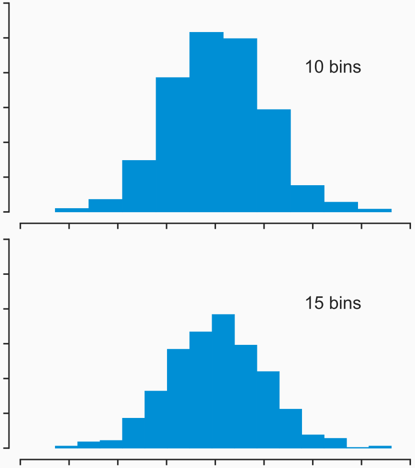
- The *histogram* is a commonly used technique for visualising the distribution of data in a sample:
    - The *x*-axis measures the *values* of the data points in the sample.
    - The *y*-axis measures the *frequency of occurrence* of the values in the sample, *i.e.* how often a given value occurs.



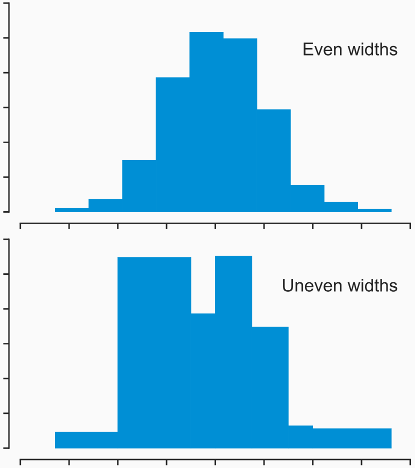
- We can create a histogram by placing the sample data points in *bins*.
- Each bin is visually represented by a vertical bar:
    - The width of the bar represents the range of the values of the sample data points contained in the bin.
    - The height of the bar represents the number of sample data points contained in the bin.



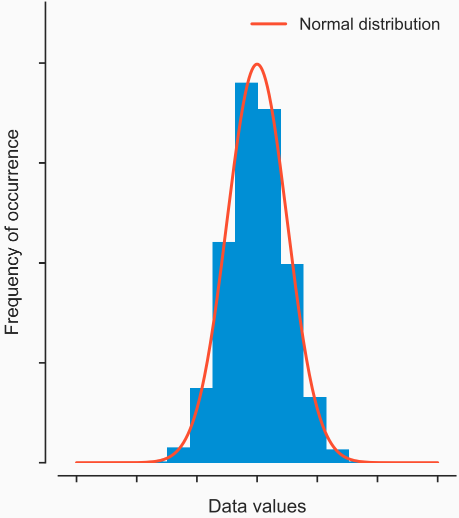
- The number of bins in a histogram is arbitrary, but the choice is important:
  - Too few bins distorts the shape of the distribution.
  - Too many bins leads to a “broken comb” look.
- The histograms on the right show the effect of varying numbers of bins on the same data sample.



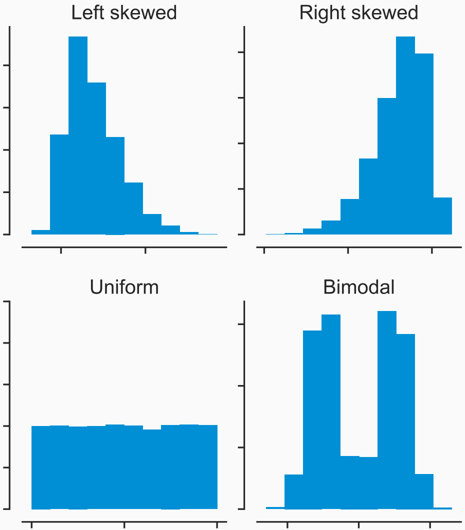
- The widths of the bins are also arbitrary, but again the choice is important:
  - Wider bins can decrease noise (spikiness) in ranges where the density of samples is low.
  - Narrower bins can increase precision in ranges where the density of data points is high.
- The histograms on the right show how uneven bin widths can distort the shape of the same distribution.



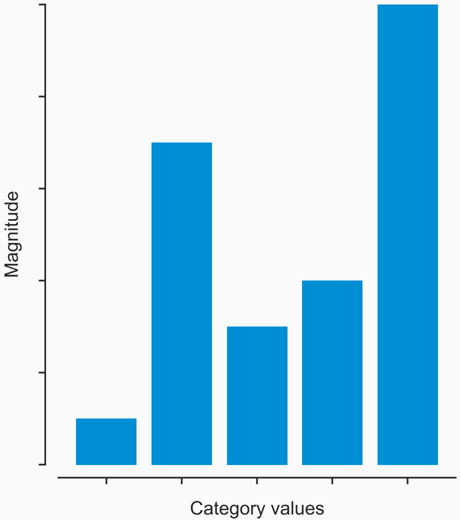
- In some situations, a distribution might look “normal”, *i.e.* it resembles the *probability density function* of the *normal distribution*.
- The normal distribution (*a.k.a.* the Gaussian distribution) is an idealised distribution with useful mathematical properties.
- In cases where a sample distribution is very close to a normal distribution, we can exploit these properties to simplify our analysis (more on this later!).



- However, not every distribution is normal!
- Sometimes, distributions have long tails - this property is known as *skewness*.
- Distributions can also be flat or *uniform*.
- Distributions with two peaks are known as *bimodal* or, more generally, *multimodal*.
- It's important to understand the shape of your distributions before conducting any complex analysis.



- In cases where we are dealing with categoric data, we cannot use a histogram to represent the distribution of values, *e.g.*
    - Numbers of cat images and dog images.
    - The proportion of spam to non-spam email.
    - The countries of origin of different customers.
- Instead, we can use a bar chart (or a pie chart) to visualise the proportions of the values in our sample.



Summary

X.1 / SUMMARY

- Lots of material this week!
  - Data types.
  - Visual cues and how to use them.
  - A process for data visualisation — why, what, how and who?
  - Distributions.
- This week's lab focuses on how to apply visual EDA techniques with pandas:
  - Bar charts.
  - Pie charts.
  - Histograms.
  - Scatter plot matrices.
- Next week, we'll look at:
  - Statistics.
  - Anomalies.
  - Relationships.

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