Imperial College London

Data Exploration & Visualisation

John Pinney

October 2019

Outline

Exploratory Data Analysis

- Distributions
- Outliers, errors, missing data
- Variation and covariation
- Asking questions about data
- Clustering

Graphics for Communication

- Principles of data graphics
- Good and bad practice
- Accessibility
- Grammar of graphics

Exploratory Data Analysis

What is exploratory data analysis?

- An approach to data analysis that focuses on summarising the main characteristics of data.
- Often employs visualisation methods.
- Can be used to help formulate hypotheses.
- May reveal unexpected patterns in the data.

Example data set

A 1987 study of different types of glass for criminological investigation.[1]

Types of glass

- 1 building windows (float processed)
- 2 building windows (non-float processed)
- 3 vehicle windows (float processed)
- 4 vehicle windows (non-float processed)
- 5 containers
- 6 tableware
- 7 headlamps

The float process for making very flat glass sheets was invented in the 1950s

Problems with Spreadsheets...

Fragile analysis: easy to introduce errors and hard to detect them.

e.g. Excel has long been known to mangle gene names! [2]

Analysis done with spreadsheets can also be difficult to reproduce.

Orange

Orange is one approach to making data analysis more robust and reproducible.

`https://orange.biolab.si/`

The application is based on a visual programming paradigm: you construct a workflow by chaining together different widgets.

Orange

Orange contains a wide range of widgets for

- data handling
- visualisation
- machine learning

Widget outputs can be assembled into reports and exported to PDF or HTML.

CSV

CSV (comma separated values) files are tables where columns are delimited by `,`

A header row gives the name for each column, e.g.

```
type,RI,Na,Mg,Al,Si,K,Ca,Ba,Fe
3,1.51655,13.41,3.39,1.28,72.64,0.52,8.65,0,0
2,1.51851,13.2,3.63,1.07,72.83,0.57,8.41,0.09,0.17
1,1.51742,13.27,3.62,1.24,73.08,0.55,8.07,0,0
1,1.52213,14.21,3.82,0.47,71.77,0.11,9.57,0,0
2,1.53125,10.73,0,2.1,69.81,0.58,13.3,3.15,0.28
```

Task

Export each lab's data from Excel to a separate CSV file.

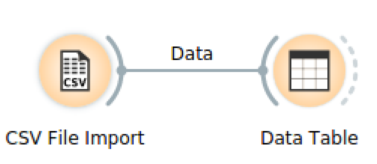
Importing CSV

Task

Use a **CSV File Import** widget to load the data for Lab 1.

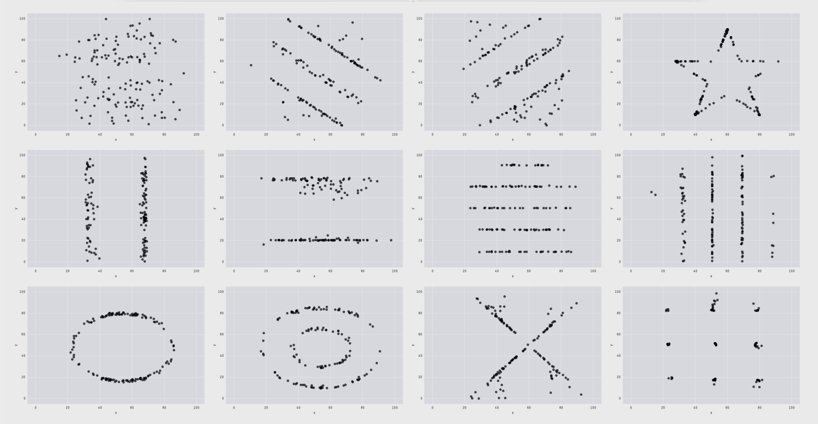
Make sure that the `type` column is treated as a categorical variable.

View the data as a **Data Table**.

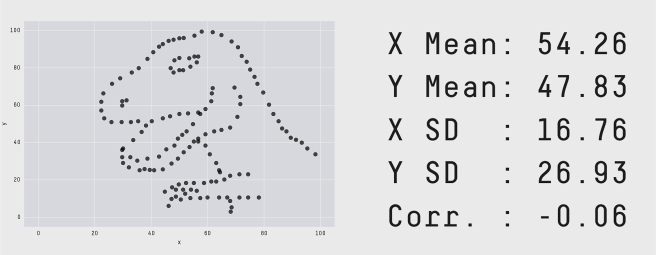


Looking at data

Q: What do these data sets have in common?



Looking at data



```
X Mean: 54.26
Y Mean: 47.83
X SD  : 16.76
Y SD  : 26.93
Corr. : -0.06
```

A: They all have the same summary statistics as the Datasaurus [3].

Always look at the data!

Distributions

Task

Use a **Distributions** widget to look at histograms for each column.

Which variables appear to be normally distributed?

What happens when the data are split by `type`?

Combining data sets

Task

Use two more **CSV File Imports** to load the data for Labs 2 and 3.

Use a **Concatenate** widget to combine data from all three labs into one data set.

Are there any differences in the data from each lab?

Use the *Report* icon to make a note of anything interesting.

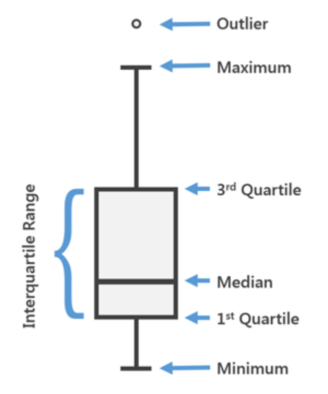
Dealing with errors

Task

Insert a **Select Rows** widget to take care of suspicious outliers.

Insert a **Select Columns** widget to eliminate any columns that appear to contain systematic errors.

Box plots



Task

Use the **Box Plot** widget to compare variable distributions between `types`.

Which variables have significantly different means between `types`?

Scatter plots

Task

Add a **Scatter Plot** widget to look at covariation between variables.

Use *Find Informative Projections* to find the pairs of variables that maximise separation between `type`s.

Use a **Correlations** widget to suggest other interesting projections of the data.

Predicting RI from the other features?

We may identify continuous-valued features (e.g. `RI`) that we would like to predict from the other features.

This is a regression modelling task, which is a type of machine learning.

Orange provides widgets to train and test linear regression models, but this is beyond the scope of today's workshop.

Imputation

Scatter plots based on features can be informative by themselves. However, to explore patterns in high-dimensional data, we usually need to find projections onto axes that are *linear combinations* of features.

To do this, we want every data point to have a value for every feature. Missing data will cause problems for most analysis methods.

Instead of dropping data points, one solution is to impute values where they are missing in the table.

Imputation

There are a variety of ways to impute missing data.

The simplest approach is to insert the mean value of the variable - this should ensure that the imputed value does not bias any downstream analysis.

However, imputation will change the distribution of data, so we should be careful about drawing strong conclusions from a data set containing imputed values.

Task

Add an **Impute** widget to deal with missing values.

PCA

Principal Component Analysis finds a set of orthogonal directions in the (normalised) data space that maximise variance.

This is an *unsupervised* analysis, as it does not depend on the data labels.

Sometimes PCA can be a helpful part of exploratory data analysis, but it is not the only way to look for patterns in the data.

PCA is *not* a clustering method.

PCA

Task

Add a **PCA** widget to calculate the first three principal components.

You will need to insert a **Select Columns** widget to set `type` as the *target* variable.

Look at the transformed data and the weights for each component.

FreeViz

FreeViz is a *supervised* approach to finding an informative 2D projection of a high-dimensional data set [4].

It adjusts the weights for each feature so as to maximise the separation between the classes of the target variable. You can also move the feature vectors manually.

FreeViz can be an intuitive way to find a subset of features that explain the differences between classes, i.e. to perform feature selection.

FreeViz

Task

Use a **FreeViz** widget to find a good projection for separating `type`s.

You should find that type 2 glass appears quite heterogeneous. Take it out of the data set temporarily by inserting a **Select Rows** widget.

Now use FreeViz to find features that can separate

- float from non-float glass
- headlamps from other non-float glass
- tableware from containers

Predicting `type` from the other features?

We have identified a few simple rules for distinguishing between some of the `type`s.

In machine learning, predicting a categorical variable from a set of features is called classification. This is a *supervised* task.

Assuming there is sufficient information in the data provided, a good classification method will find ways to distinguish classes reliably, even when they are not linearly separable in the original feature space.

Orange provides widgets to train and test a variety of classification models, but this is beyond the scope of today's workshop.

k-means clustering

Clustering is an *unsupervised* task that tries to divide the dataset into subgroups that contain similar data points. It can be a useful element of exploratory data analysis.

k-Means clustering operates directly on the feature space. This means that we do not need to compute distances between data points.

k-means is a *heuristic* method, which means that it may not always produce the same result.

k-means clustering

Task

Use a **k-Means** widget to find clusters in the data set.

How many clusters appear to be present?

Which features can be used to separate clusters?

How do the clusters correspond to `type`?

Your turn

Explore your own data

Task

Apply these exploratory data analysis techniques to your own tabular data set and prepare a report on your findings.

No suitable data of your own? The **Datasets** widget has lots of examples to play with.

Choose a data set that

- has at least 5 variables.
- has more instances than variables.
- is not tagged *synthetic* or *image analytics*.

Explore your own data

Tips

Columns must correspond to variables and need variable names as a header row.

Categorical variables need to be identified during CSV import.

Use **Data Table** and **Distributions** for an initial sanity check.

Use **Select Rows** to work with a subset of the data.

Use **Select Columns** to specify the target variable.

Use **Impute** to fill in missing data if needed.

Graphics for Communication

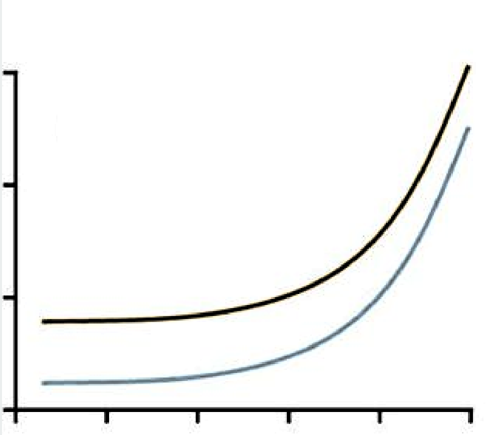
Visual estimations

Human perception treats different types of visual stimuli in different ways.

Some types of representation are easier to decode and compare than others.

Our tendency to visually assemble elements that are close together or similar to each other can also cause biases in perception that can distort the message communicated. [5]

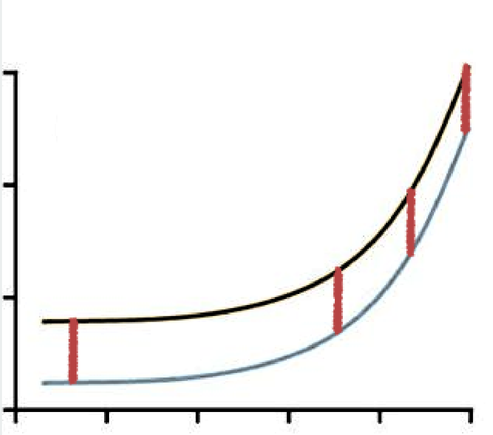
Visual estimations



This plot shows two variables that increase over time.

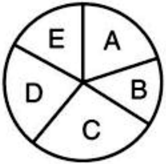
What proportion of the initial difference (y_1 - y_2) is the final difference?

Visual estimations



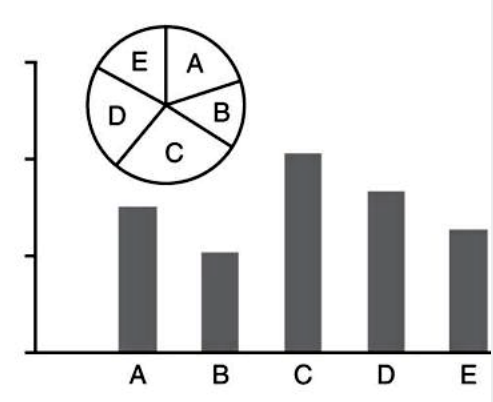
In fact, there is no change in the difference between the curves.

Visual estimations



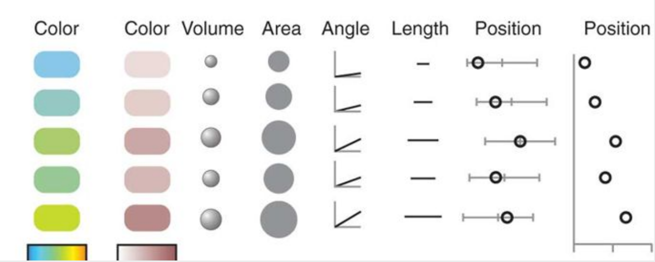
Order the letters by decreasing value.

Visual estimations



Comparing pie chart areas is notoriously difficult.
A bar chart is much easier to decode.

Visual estimations



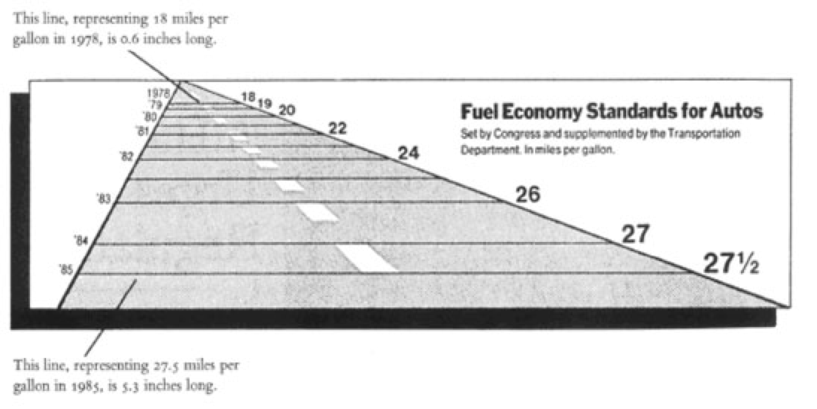
The data in each column are the same, but the perceived values may be quite different.

Principles of data graphics

Edward Tufte presents the following principles for visual communication. [6]

- Tell the truth
  Show the data and avoid distorting what they have to say.
- Show data variation
  Lead the viewer to focus on the substance of the findings rather than the methodology or graphical design.
- Make large data sets coherent
  Present data efficiently and encourage the eye to compare different pieces of data.
- Reveal the data at several levels of detail
  From a broad overview to the fine structure.

Visual Integrity



This line, representing 18 miles per gallon in 1978, is 0.6 inches long.

This line, representing 27.5 miles per gallon in 1985, is 5.3 inches long.

Visual Integrity

It is easy to design graphics that deliberately mislead the viewer. We can quantify the amount of distortion using Tufte's lie factor:

$$\text{lie factor} = \frac{\text{size of effect shown in graphic}}{\text{size of effect in data}}$$

where

$$\text{size of effect} = \frac{|\text{second value} - \text{first value}|}{\text{first value}}$$

The lie factor of the fuel economy graph is 14.8.

(numerical change = 53% but graphical change = 783%)

Data-Ink

Good graphical representations should maximize data-ink and erase as much non-data-ink as possible.

The data-ink ratio is defined as

$$1 - \text{proportion of ink that can be erased without loss of information}$$

Data-Ink

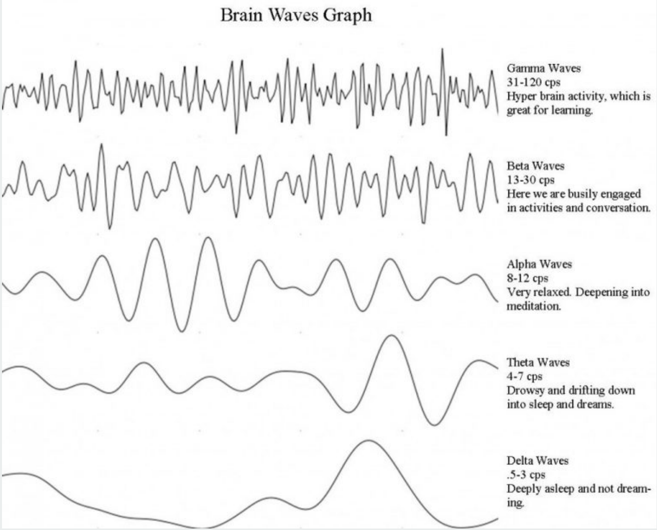
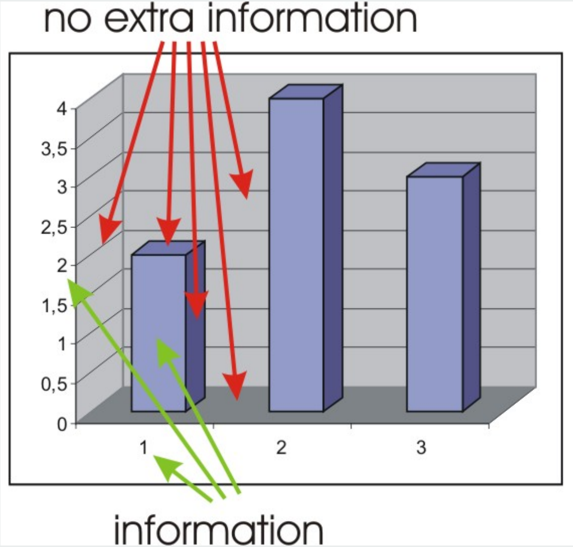

Brain Waves Graph
Gamma Waves
31-120 cps
Hyper brain activity, which is great for learning.
Beta Waves
13-30 cps
Here we are busily engaged in activities and conversation.
Alpha Waves
8-12 cps
Very relaxed. Deepening into meditation.
Theta Waves
4-7 cps
Drowsy and drifting down into sleep and dreams.
Delta Waves
.5-3 cps
Deeply asleep and not dreaming.

Chart Junk

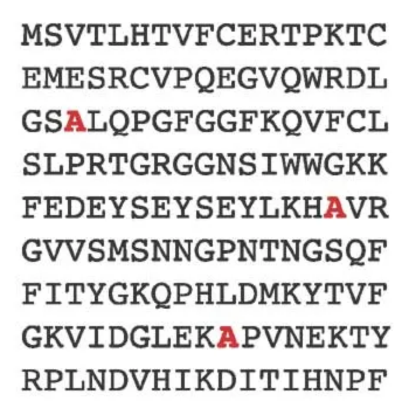
”The interior decoration of graphics generates a lot of ink which does not tell the viewer anything new. The purpose of the decoration varies - to make the graphic appear more scientific, to enliven the display, to give the designer an opportunity to exercise artistic skill. Regardless of the cause, it is all non-data-ink or redundant data-ink, and it is often chartjunk.”[6]

Chart Junk



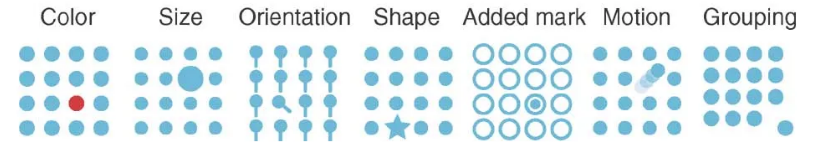
Salience

The visual salience of a graphical element is the extent to which it stands out from its surroundings.[7]



MSVTLHTVFCERTPKTC
EMESRCVPQEGVQWRDL
GSALQPGFGGFKQVFCL
SLPRTGRGGNSIWWGKK
FEDEYSEYSEYLKHAVR
GVVSMSNNGPNTNGSQF
FITYGKQPHLDMKYTVF
GKVIDGLEKAPVNEKTY
RPLNDVHIKDITIHNPF

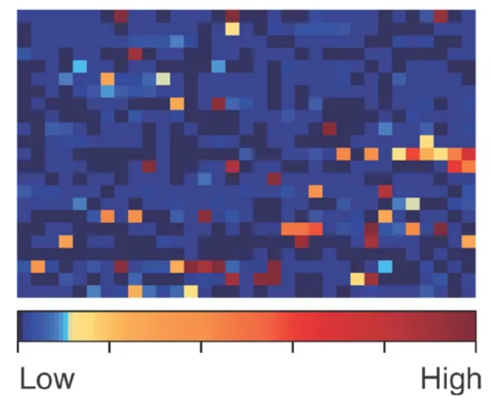
Salience



Examples of visual features that make objects distinct.

Salience vs relevance

It is important that the features that are most noticeable (high salience) are those that are most important to the message communicated (high relevance).[8]

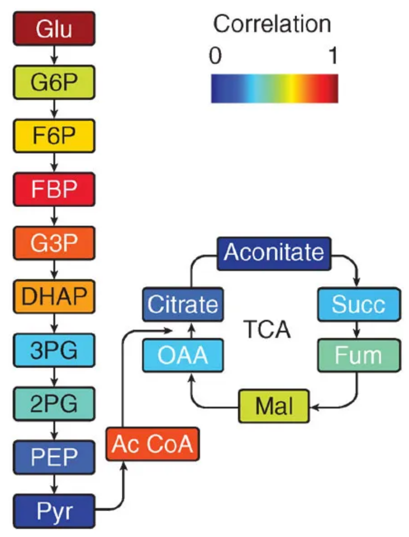


Use of colour

Colour can be very helpful for distinguishing between classes (categorical data).

It can be more difficult to construct an effective visualisation for quantitative data using a colour scale, because perceived change in hue / saturation / brightness do not map evenly to values [9].

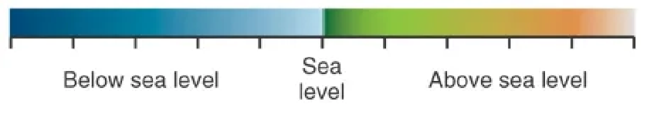
Use of colour



Rainbow scales have particularly uneven transitions and can be difficult to interpret.

Use of colour

Changes in hue have very high salience, so can be useful to emphasise zero-crossings where appropriate (e.g. a topographical map)



In general it will help communication to avoid relying on colour to communicate information [10].

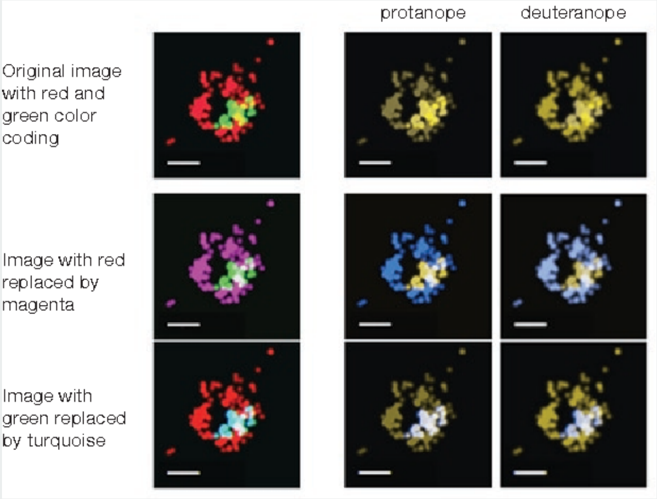
Colour blindness

Problems with colour vision are very common - around 4.5% of people in the UK are affected.

Any audience bigger than 15 is *more likely than not* to include a colour-blind person.

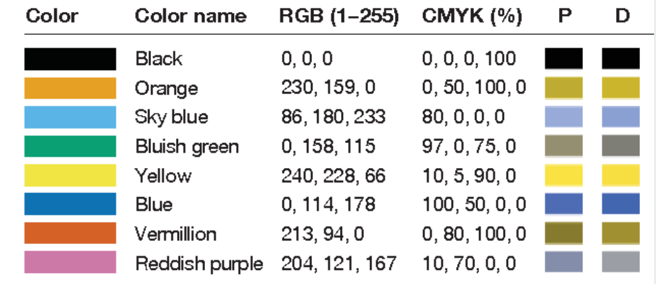
It is easy to adjust colours to avoid coding information as red/green (the most common form of colour blindness). [11]

Colour blindness


protanope
deuteranope
Original image with red and green color coding
Image with red replaced by magenta
Image with green replaced by turquoise

Colour blindness

If several colours are needed, use a palette that maximises the distinction between different colours.



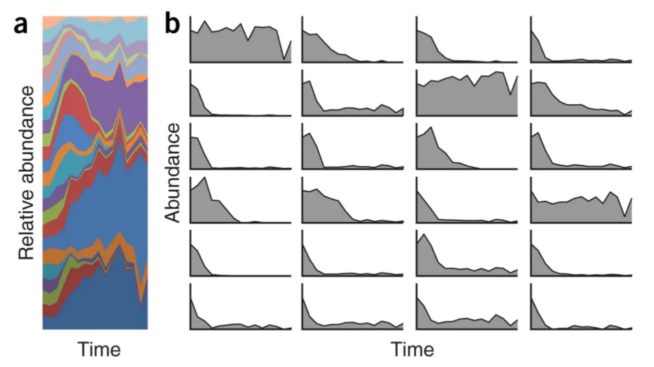
Color	Color name	RGB (1–255)	CMYK (%)	P	D
⬛	Black	0, 0, 0	0, 0, 0, 100	⬛	⬛
🟧	Orange	230, 159, 0	0, 50, 100, 0	🟨	🟨
🟦	Sky blue	86, 180, 233	80, 0, 0, 0	🟦	🟦
🟩	Bluish green	0, 158, 115	97, 0, 75, 0	🟫	⬜
🟨	Yellow	240, 228, 66	10, 5, 90, 0	🟨	🟨
🟦	Blue	0, 114, 178	100, 50, 0, 0	🟦	🟦
🟥	Vermillion	213, 94, 0	0, 80, 100, 0	🟫	🟫
🟪	Reddish purple	204, 121, 167	10, 70, 0, 0	⬜	⬜

Small multiples

Presenting *classes of behaviour* in high-dimensional data is a common visualisation goal.

By preparing lower-dimensional *slices* of the data, we can draw attention to the similarities and differences in behaviours [12].

Small multiples

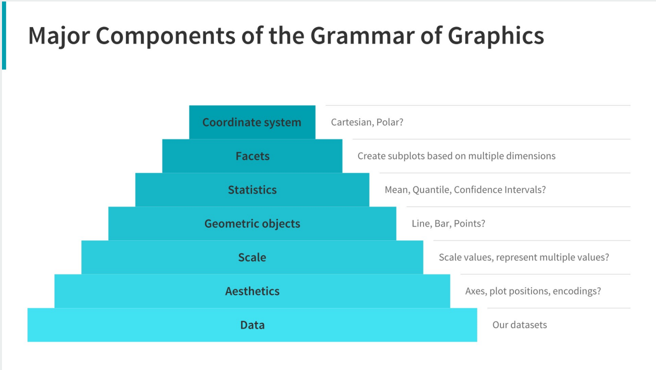


Grammar of graphics

A grammar of graphics is a way to conceptualise the components of a graphic. It helps us to move beyond standard plots (e.g. ”the scatterplot”) and combine visual elements in ways that best suit the information to be communicated.

Originally proposed by Leland Wilkinson and developed by Hadley Wickham [13].

Grammar of graphics



Major Components of the Grammar of Graphics

Coordinate system	Cartesian, Polar?
Facets	Create subplots based on multiple dimensions
Statistics	Mean, Quantile, Confidence Intervals?
Geometric objects	Line, Bar, Points?
Scale	Scale values, represent multiple values?
Aesthetics	Axes, plot positions, encodings?
Data	Our datasets

Software packages for effective graphics

Grammar of graphics

- R: ggplot2
- Python: plotnine

Interactive plots

- R: shiny
- Python: plotly, bokeh

Further reading

Since 2013, Nature Methods has run a regular feature looking at data visualisation best practice, called *Points of View*. [14].

These articles cover the most common data types and can be a great source of inspiration for difficult visualisation problems.

References i

[1] https://archive.ics.uci.edu/ml/datasets/Glass+Identification

[2] Ziemann M, Eren Y El-Osta A. Gene name errors are widespread in the scientific literature. Genome Biol 17:177 (2016)

[3] http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html

[4] Demšar J, Leban G Zupan B. FreeViz—An intelligent multivariate visualization approach to explorative analysis of biomedical data. J Biomed Inform 40:661-671 (2007)

[5] Wong, B. Design of data figures. Nat Methods 7, 665 (2010)

References ii

[6] E. R. Tufte. The Visual Display of Quantitative Information, 2nd Edition. Graphics Press, Cheshire, Connecticut, 2001.

[7] Wong, B. Salience. Nat Methods 7, 773 (2010)

[8] Wong, B. Salience to relevance. Nat Methods 8, 889 (2011)

[9] Gehlenborg, N., Wong, B. Mapping quantitative data to color. Nat Methods 9, 769 (2012)

[10] Wong, B. Avoiding color. Nat Methods 8, 525 (2011)

[11] Wong B. Points of view: Color blindness. Nat Methods 8:441 (2011)

References iii

[12] Shoresh, N., Wong, B. Data exploration. Nat Methods 9, 5 (2012)

[13] H Wickham. A layered grammar of graphics. Journal of Computational and Graphical Statistics, vol. 19, no. 1, pp. 3–28, 2010.

[14] `http://blogs.nature.com/methagora/2013/07/data-visualization-points-of-view.html`