Information Visualization

Preliminary concepts

Lesson 2

Marilena Daquino
Assistant Professor

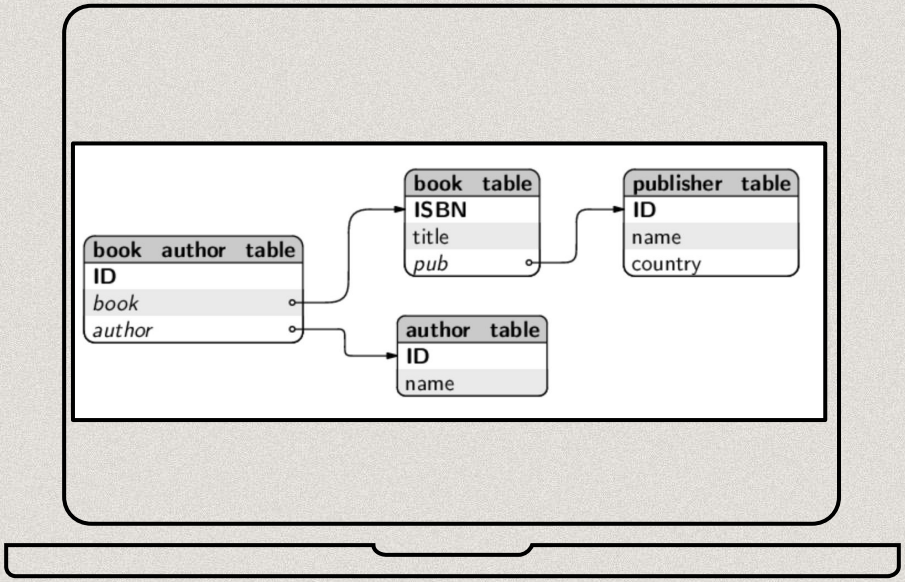
Department of Classical Philology and Italian Studies

marilena.daquino2@unibo.it

Table of contents

01

Data attributes



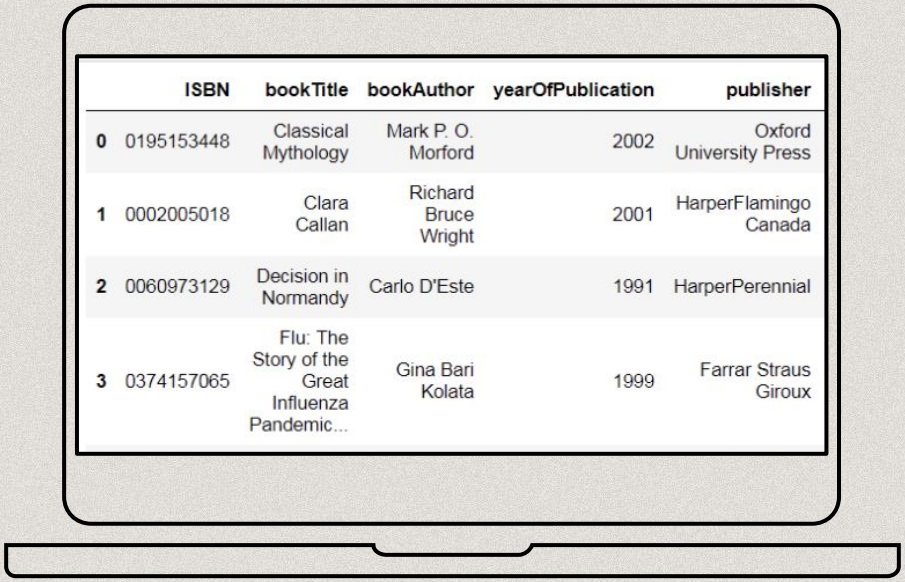
Data

Data are abstractions of concepts and real-world **entities** (people, books, places)

Data present **attributes**, also called features or variables (title, author, ISBN are attributes of books)

01

Data attributes



	ISBN	bookTitle	bookAuthor	yearOfPublication	publisher
0	0195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press
1	0002005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada
2	0060973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial
3	0374157065	Flu: The Story of the Great Influenza Pandemic...	Gina Bari Kolata	1999	Farrar Straus Giroux

Dataset

A dataset is a **collection** of entities with their attributes.

It can be seen as an **n*m matrix,** wherein n is the number of rows (entities) and m is the number of attributes (columns)

01

Data types

Attributes

Numeric

Measurable **quantities** (e.g. temperature, price)

Categorical

[also called Nominal] **Names** from non-overlapping sets, classes, or states (e.g. title, sex)

Ordinal

Nominal or numeric attributes that can be **ranked** (e.g. days, years, Likert: “strongly like” to “strongly dislike”).

Data types

Attributes

Numeric

Measurable **quantities** (e.g. temperature, price)

Interval attributes are measured on the basis of a scale with an interval and an origin (e.g. time, temperature). They have no true-zero.

Ratio attributes have a true-zero - meaning no quantity is measured in that point (e.g. price)

01

Operations on data types

Operations

Numeric

Interval can be **ordered**, but not multiplied/divided.

On Ratio you can do **arithmetic** operations.

Categorical

Can be **sorted** (e.g. alphabetically).

Cannot be ordered and arithmetic operations cannot be performed.

Ordinal

Can be naturally **ordered**.

Arithmetic operations are not possible.

01

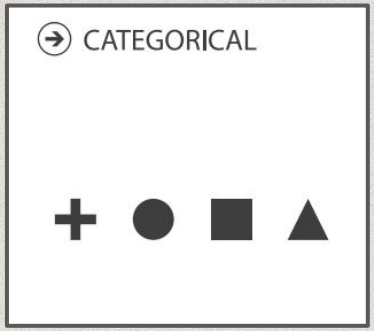
Operations on data types

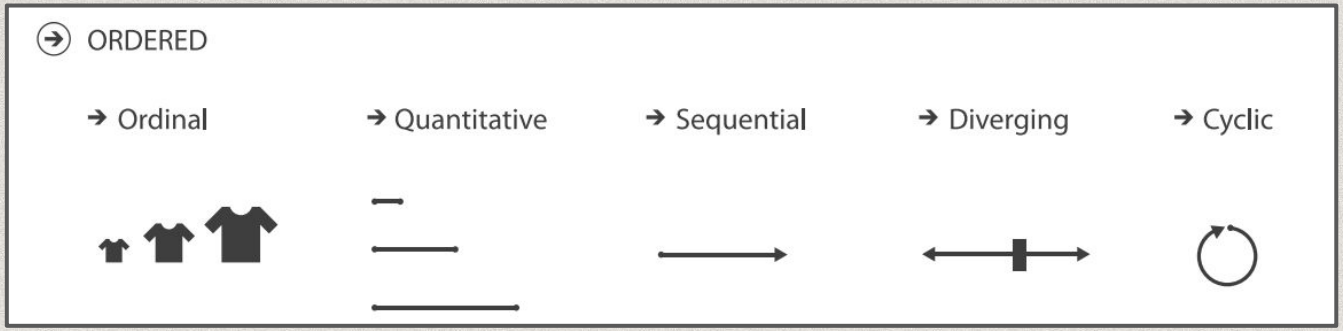
Operations

Numeric

Categorical

Ordinal





Can be ordered according to their magnitude, quantity, order in a sequence, etc.

01

Analysis on data types

Operations

Numeric

Statistical analysis

Categorical

Counting (distribution), proportions

Ordinal

Counting (distribution), proportions

02

Data viz. dimensions

A bit of history, properties and graphical elements of data visualizations, viewer’s perception

Statistical graphics

The beginning

At the end of the 18th century, William Playfair invented **statistical graphics**.

He figured that visualisations can be a fundamental aid in all research phases: **exploring, understanding, presenting**.

He was aware that visualisations cannot give precise answers (low precision in comparisons). Rather, they help to **frame** phenomena.

Statistical graphics

He designed:

- **line** charts and **area** charts (time-series data),
- **bar** charts (quantitative or categorical comparison), and
- **pie** charts (proportions within a set)

Graphical properties

Properties of a data visualization that help making noticeable graphical elements

Axes

Layout

Shape

Colour

Size

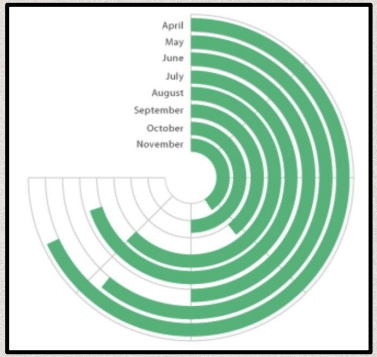
Typography

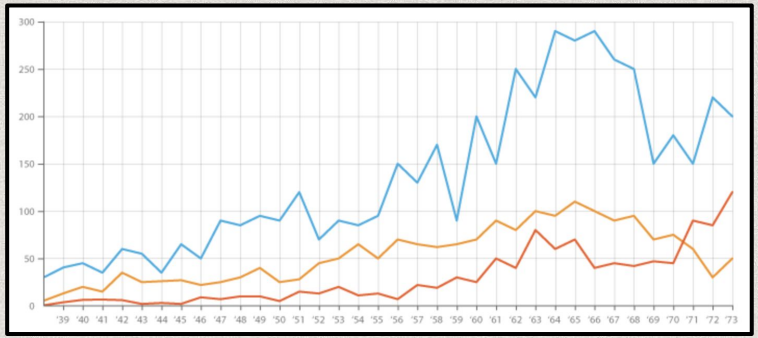
Axes

Cartesian/radial axes

A visual guide for the **placement** of elements composing the visualisation.

A visualization with axes is called **graph**, otherwise it is called chart (although chart is used for any kind of visual).





Layout

02

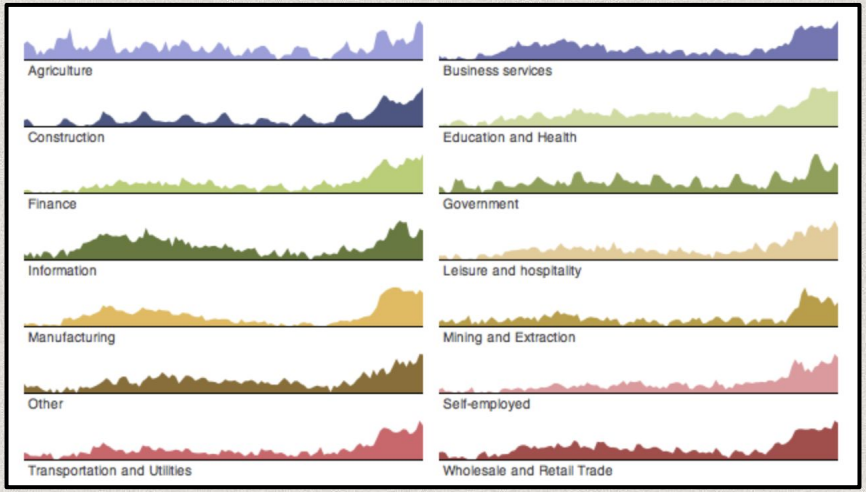
Graphical properties

Single/Multiple canvas

The **format and symmetry** of the visualisation change according to the volume of data and the number of attributes to show.

More visualisations in the canvas help comparison, but hide the big picture!

A viz. should always fit the frame (e.g. a screen).



Shape

02

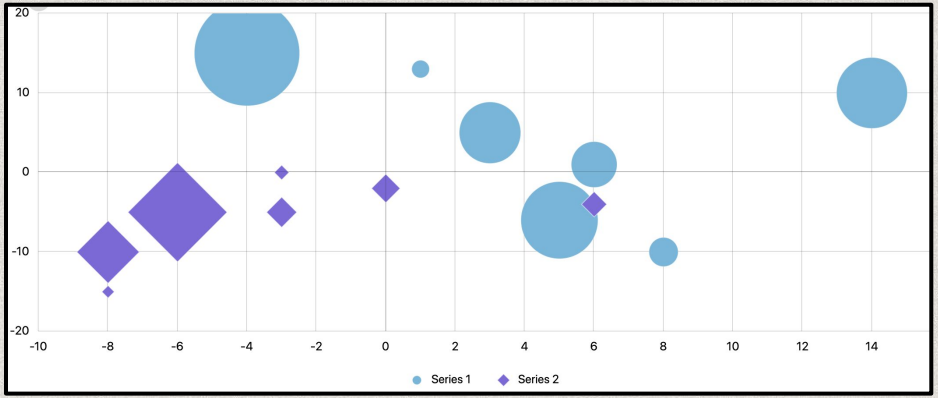
Graphical properties

Shapes and glyphs

Realistic (icons) or abstract (circles) help to **distinguish** elements.

Shapes and glyphs demand more **attention** (icons can mitigate the effort)

Lines are effective shapes (highlight a trend)



Colour

02

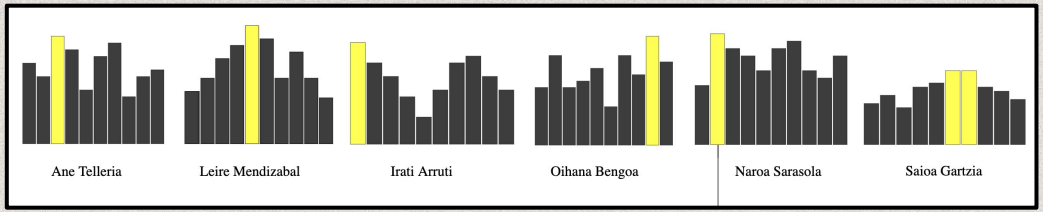
Graphical properties

Palette and luminosity

Color differences are detected in <200 milliseconds (preattentive perception). Must be natural (e.g. found in nature).

Used to **distinguish elements or patterns** in big datasets, it’s less useful in small datasets (one color is sufficient).

Luminosity highlights relevant patterns (bright colors pop out, while dark colors recede).

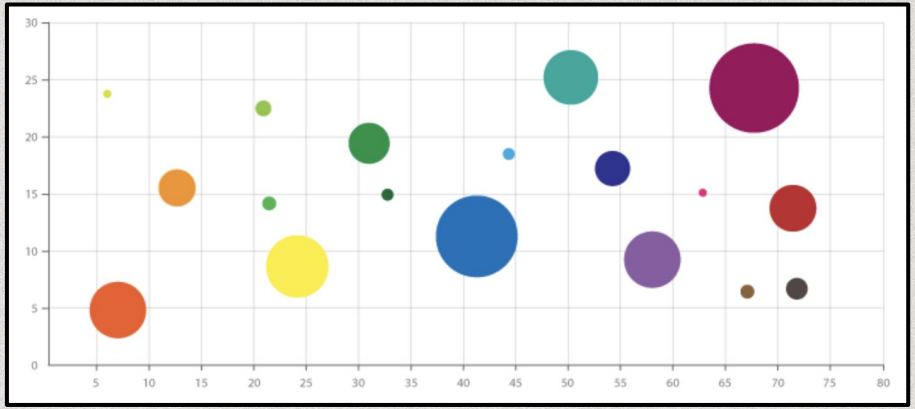


Size

Show the variety

Identify **variations between similar elements** by size (regardless of other dimensions) is the quickest way.

Still, it is not the most effective means to compare values (e.g. circle areas).



Typography

02

Graphical properties

Primary/secondary role

Can be a primary aspect (e.g. wordle) along with other dimensions (e.g. size) or secondary, to describe the visualization.

Interaction and context

02

Properties interaction

Graphical properties interact with each other to convey the message in different contexts.

Location

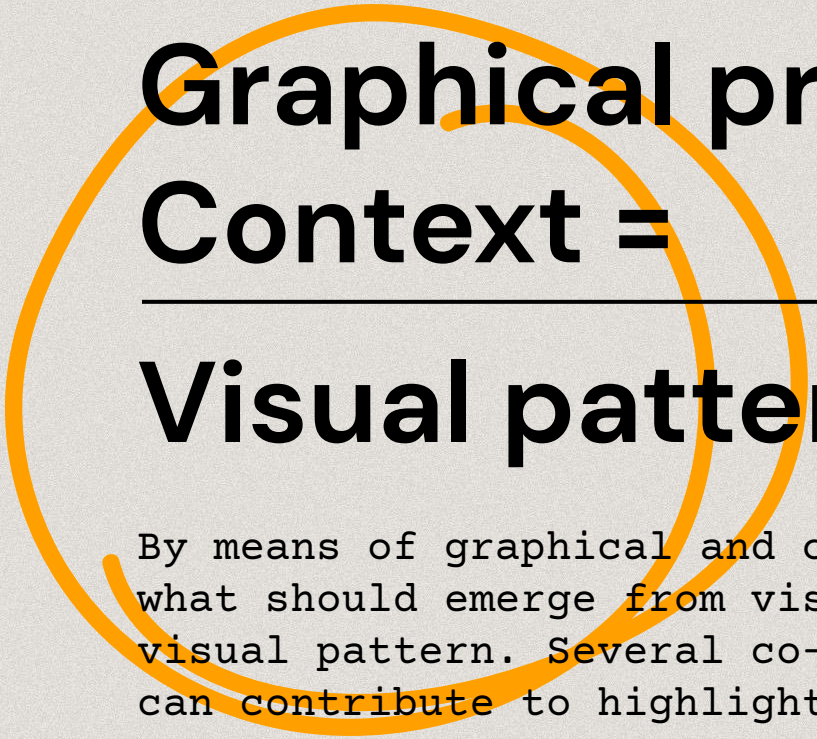
Placement helps to familiarise with data (e.g. a map)

Network

Closeness and **connections** of data points (e.g. a network graph)

Sequence

Axis-based visualisations create a **linear** reading (e.g. line graphs)



Graphical properties + Context =

Visual pattern

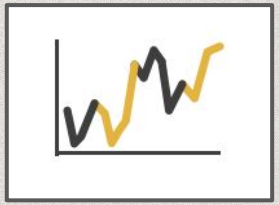
By means of graphical and contextual aspects, what should emerge from visualisations is a visual pattern. Several co-occurring dimensions can contribute to highlight patterns.

How do visual patterns look like

Differentiation

Contrast values

distribution



samples

similarity

How do visual patterns look like

02

Visual patterns

Gradation

Continuous values

trend

How do visual patterns look like

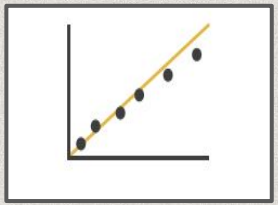
02

Visual patterns

Anomaly

Break the **pattern.**

outliers



correlation

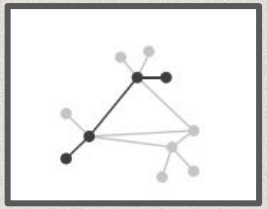
How do visual patterns look like

02

Visual patterns

Paths

Position and relations



Clusters and paths

Visual thinking

02

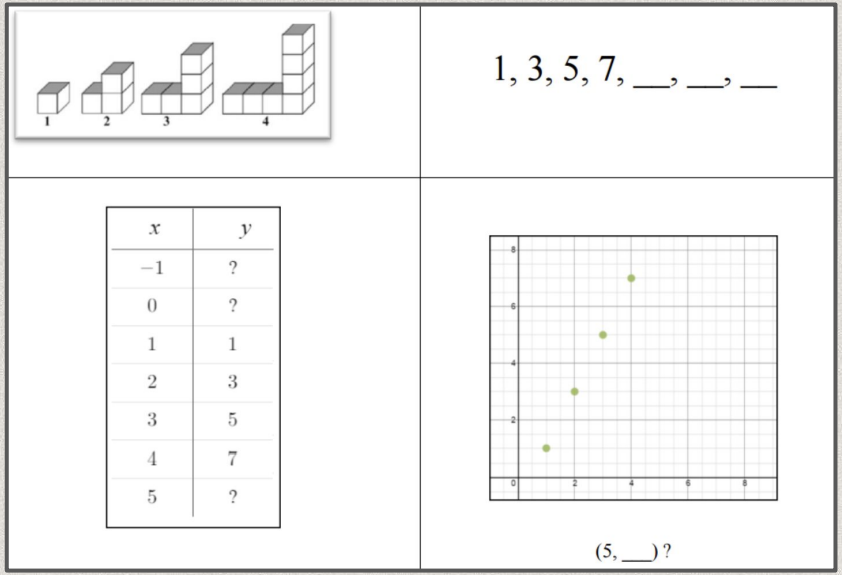
Visual patterns

Visual patterns stimulate types of thinking that cannot be achieved by looking at numbers only (top-right)

Relational thinking is when visual aids can be compared (top-left).

Recursive thinking is when you understand how many cubes to add to the next (5th) L-shape.

Functional thinking is when you know how to compute the number of cubes of the, say, 10th L-shape without drawing.



Perception

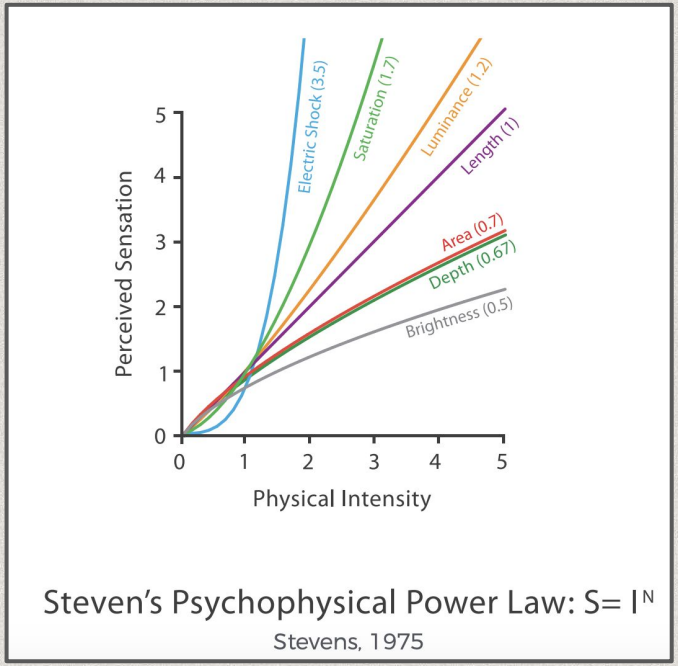
Preattentive variables

Visual aspects have a **functional** role in the interpretation of data rather than aesthetic.

The idea of visualising data in a graphical form is to **replace cognition with perception**.

The human eye perceives visual dimensions in different ways, with more or less effort, and users' **knowledge background** can affect the perception.

Nonetheless, some dimensions are more **intense** than others.

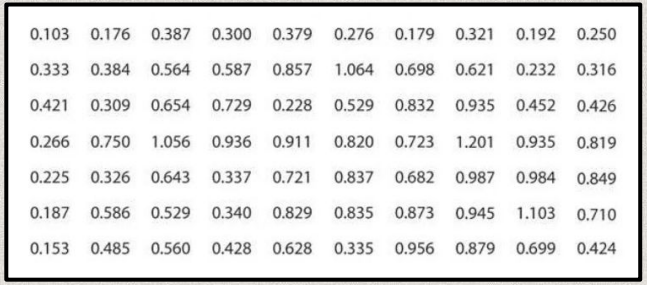


Steven's power law

Tidwell's preattentive variables

Preattentive variables

Some visual aspects work **preattentively**, meaning they are able to communicate something before the user pays conscious attention to it.



0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Find numbers greater than 1

Tidwell's preattentive variables

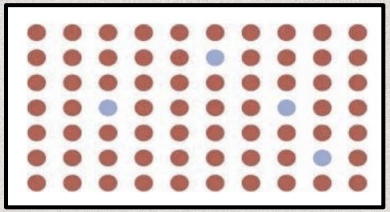
Preattentive variables

Effective and impactful data visualisations work extensively on these aspects.

Tidwell et al, Designing interfaces found 8 variables.

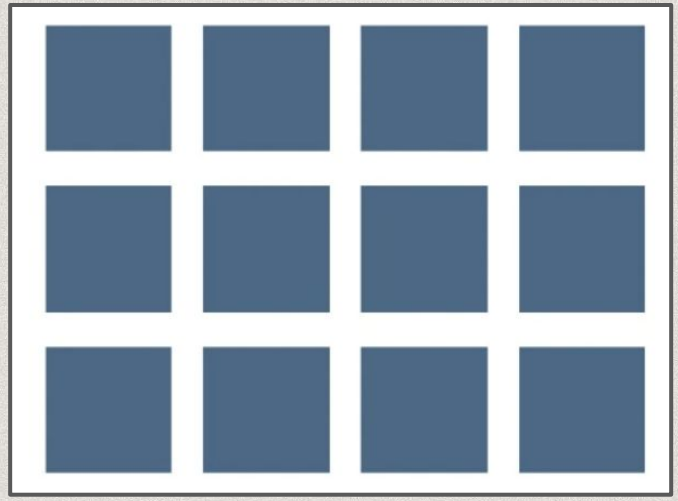
0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Find numbers greater than 1



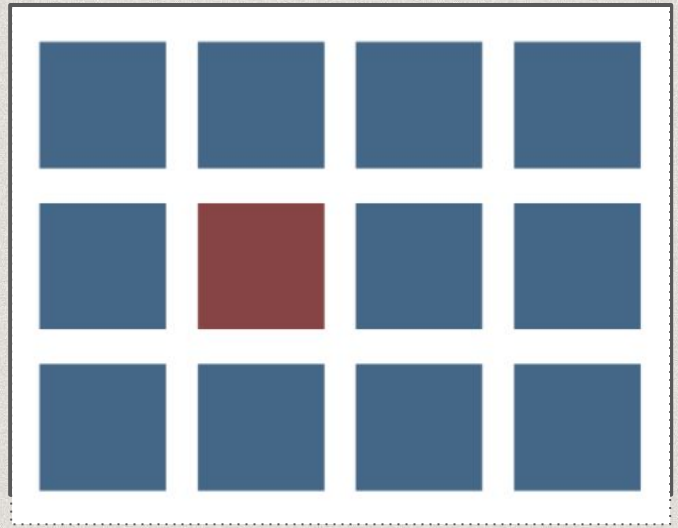
0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Tidwell's preattentive variables



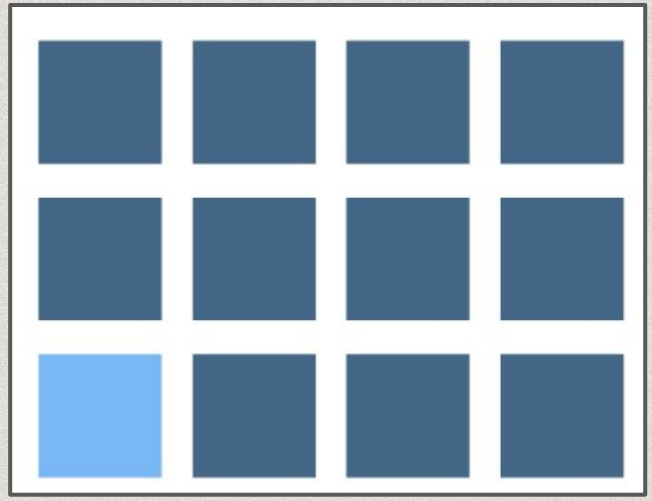
Let's start!

Tidwell’s preattentive variables



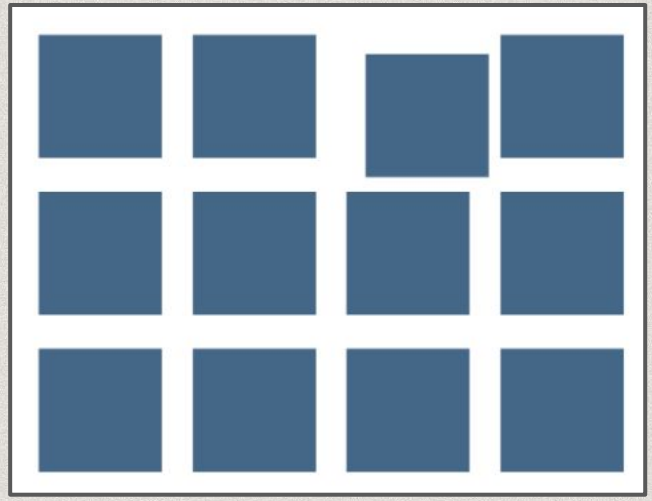
Color hue

Tidwell’s preattentive variables



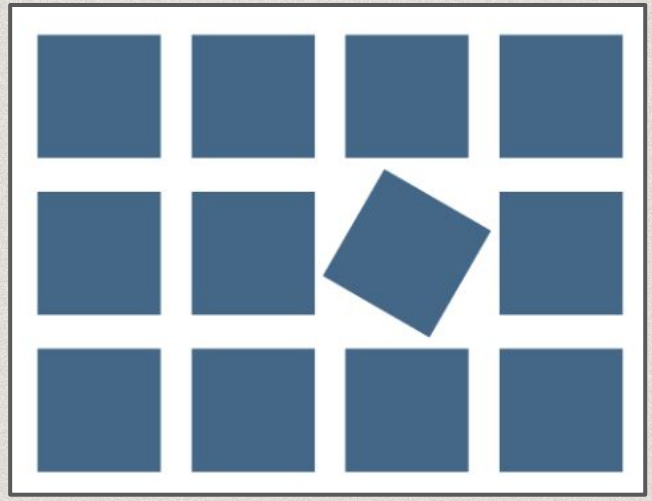
Color brightness

Tidwell's preattentive variables



Position and alignment

Tidwell's preattentive variables

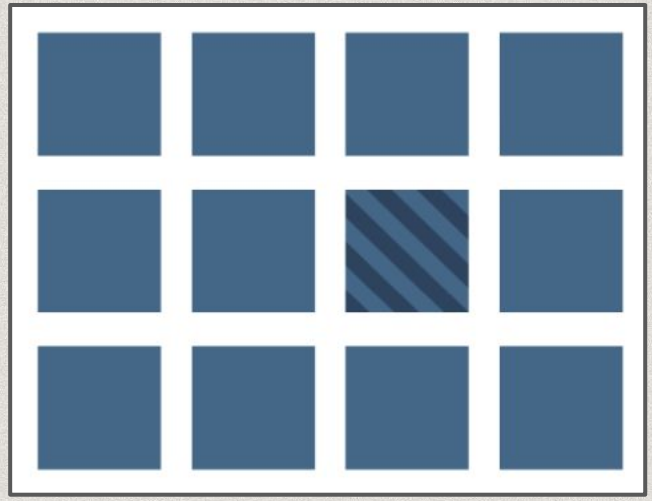


Orientation

Tidwell’s preattentive variables

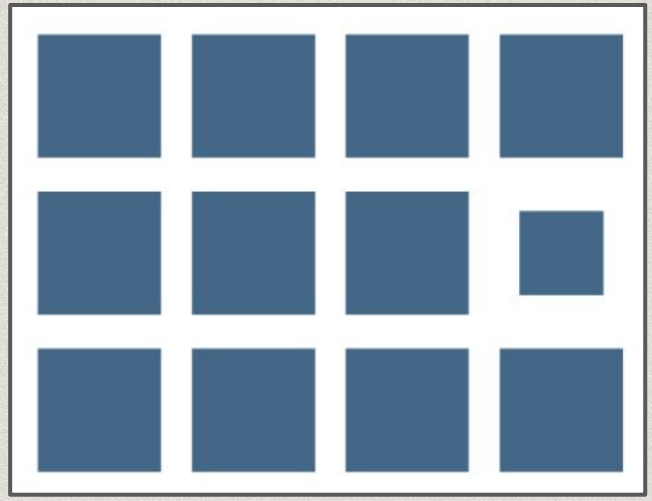
Color saturation

Tidwell’s preattentive variables



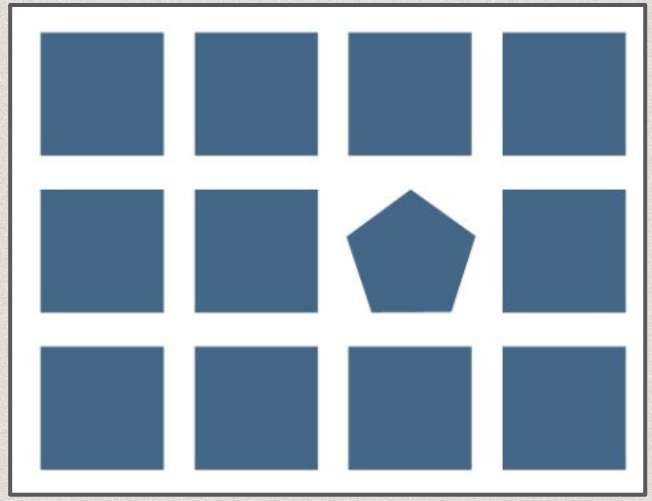
Texture

Tidwell’s preattentive variables



Size

Tidwell's preattentive variables



Shape

03

Techniques

Expectations and best practices

Expectations

Effective visualisations are able to identify, **summarise** and prioritise information.

The visual summary must be able to:

- **explore** a dataset
- **discover** knowledge

Exploratory Data Analysis

Objectives

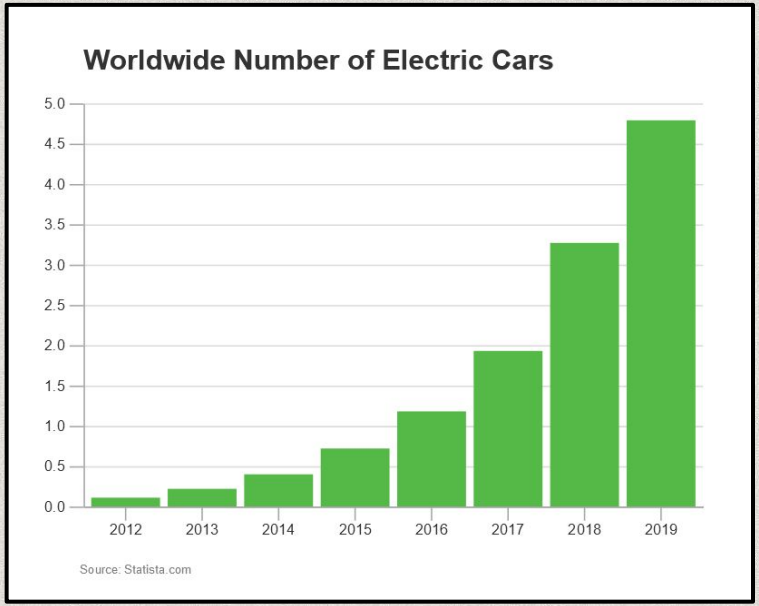
When we do not know what is inside a dataset and we need to figure out what its **added value** is (what is good for, which types of analyses it enables)

Knowledge discovery

It's **hypothesis-driven analysis**.

The designer has a question to answer, and a hypothesis and wants to demonstrate.

The validation consists of the **visual evidence** of a trend, a behaviour, or a relationship.



EDA and Knowledge discovery

03

Objectives

Exploration can be seen as a form of discovery.

Select data that may characterise the dataset (the **added value**) and find what data are available, significant, interesting and at scale for further analysis.

Again expectations

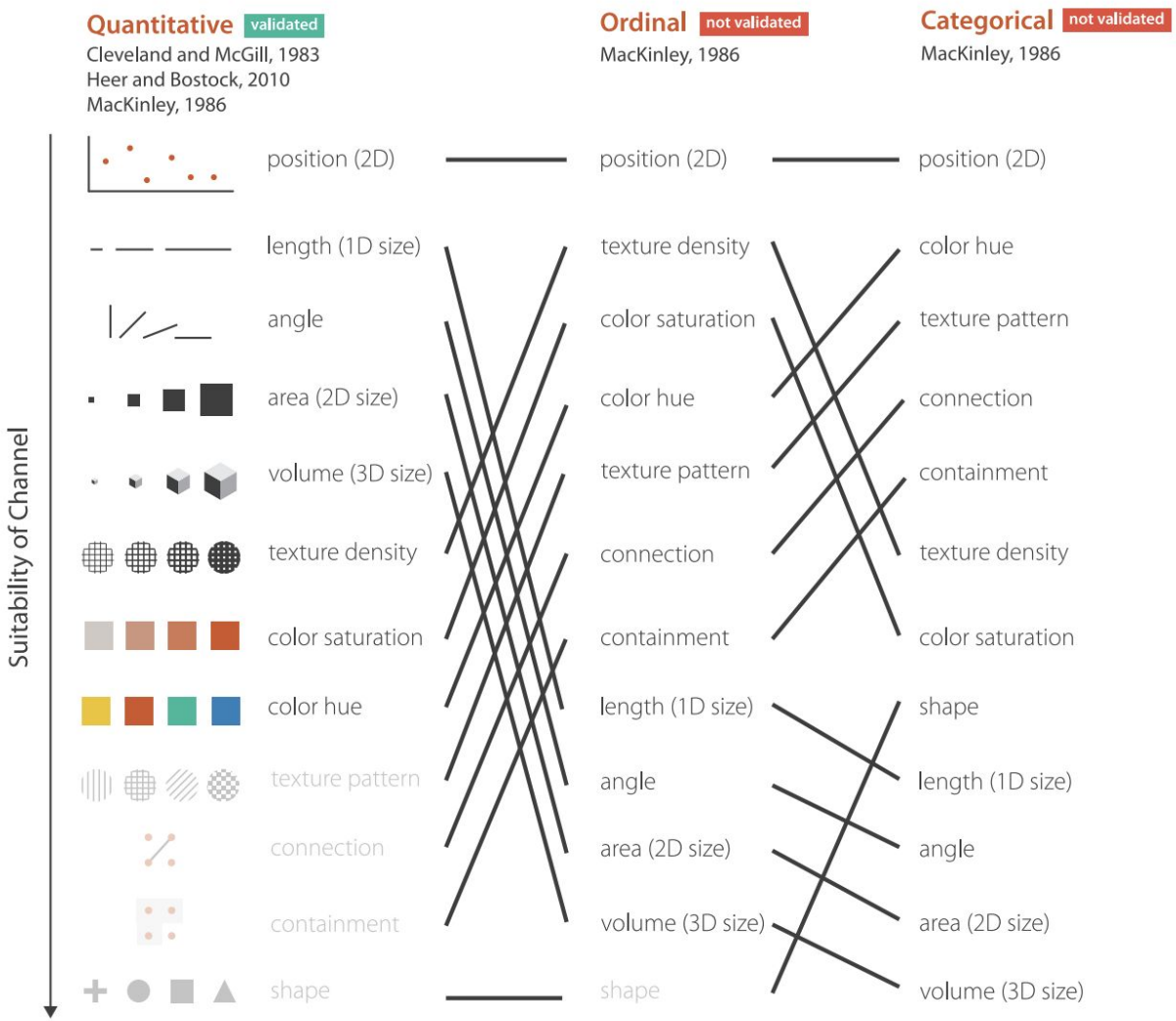
Expected positive	Unexpected negative
Expected negative	Unexpected positive

Outcomes of data analysis/discovery techniques fall into two categories:
expected/unexpected
positive/negative

Again expectations

Expected positive	Unexpected negative
Expected negative	Unexpected positive

Unexpected negative results are the most insightful patterns to support decision-making

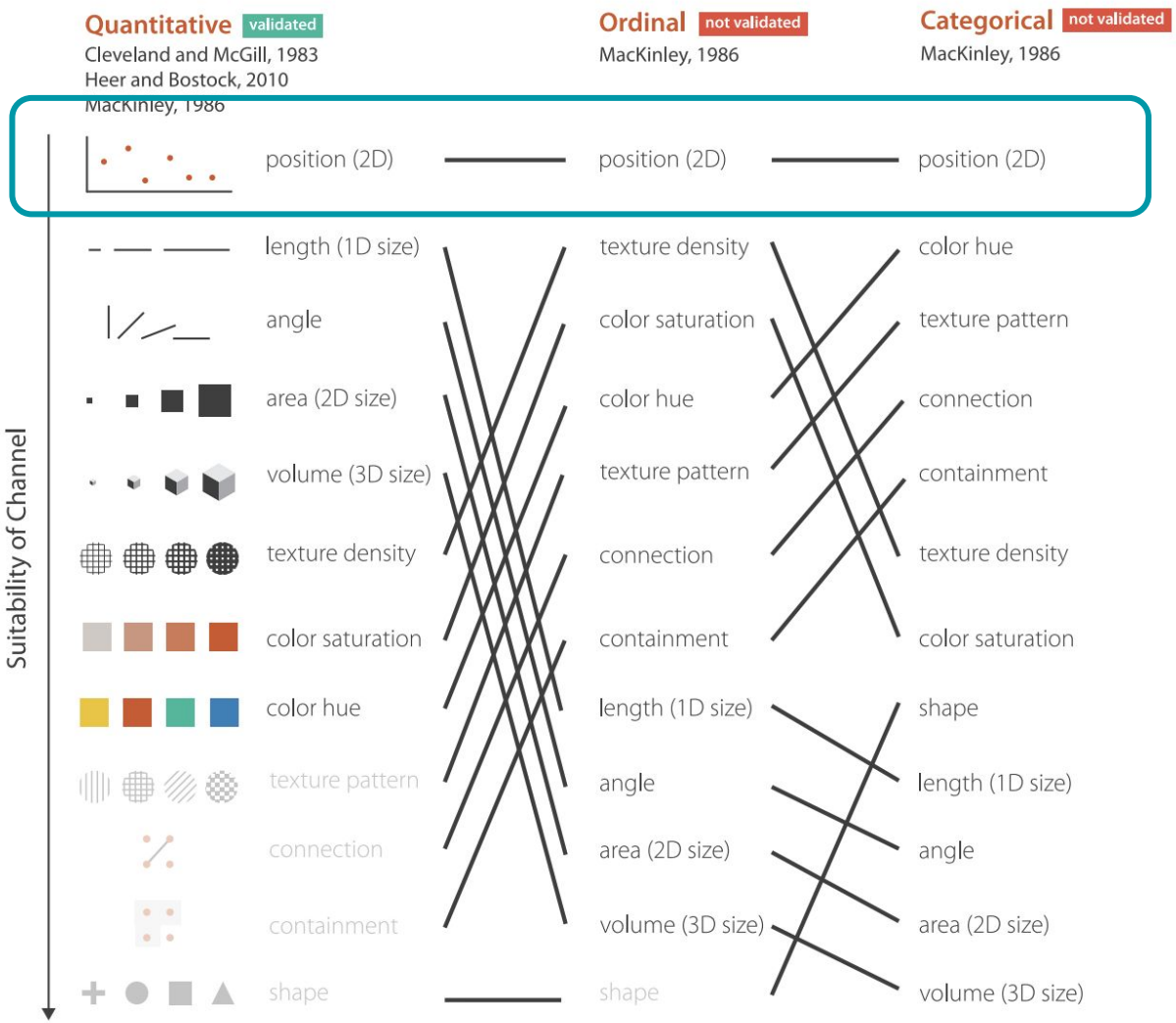


03

Best practices

Studies demonstrated that according to the **type of data attributes** at hand, certain **graphical properties** work better than others.

In particular, properties are perceived with more or less **accuracy**.

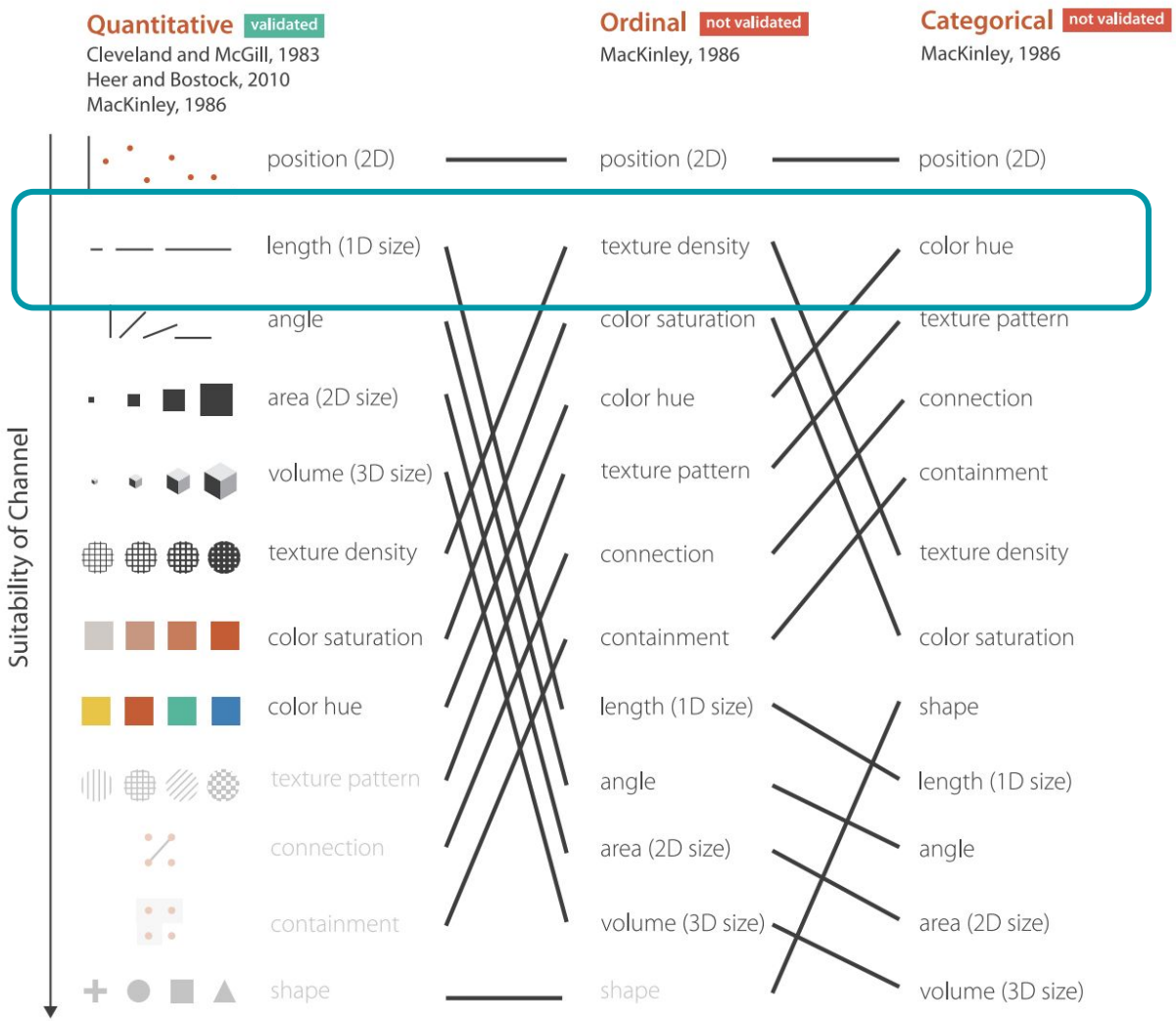


03

Best practices

Relative location (position) of elements using the same scale is the most effective property regardless of the data type

(maps, axes…)



03

Best practices

Length works better than angles (ber chart VS pie chart) when comparing numeric values.

Colors help to discriminate non quantitative values, as well as texture for ordinal values.

Edward Tufte's principles

There are several aspects that designer can tweak to affect the perception of visuals.

That said, inaccuracy and sloppiness may have unexpected results.

Edward Tufte designed some good practices to be taken into account when **finalising a visualization**.

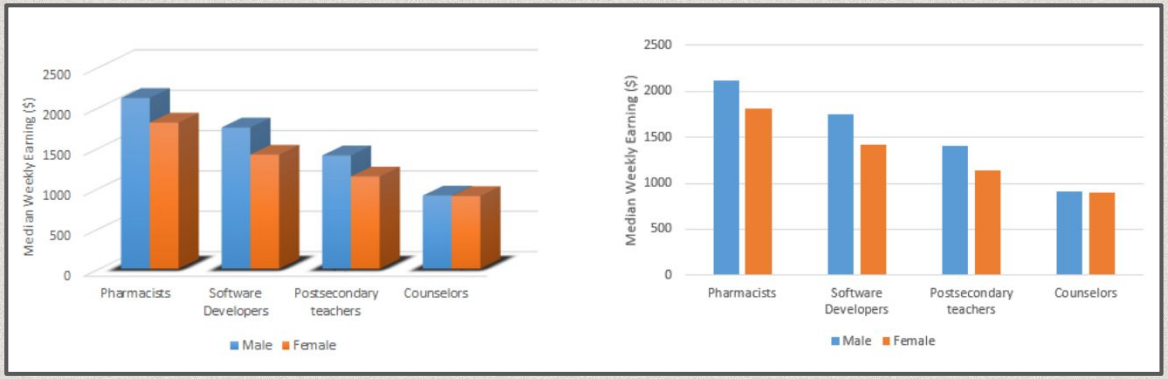
Skiena S. Data Science Design Manual. Springer. 2017

Edward Tufte's principles

Best practices

Minimize ink-ratio

The ink used to represent data should be minimal, so that the intended message is clear.

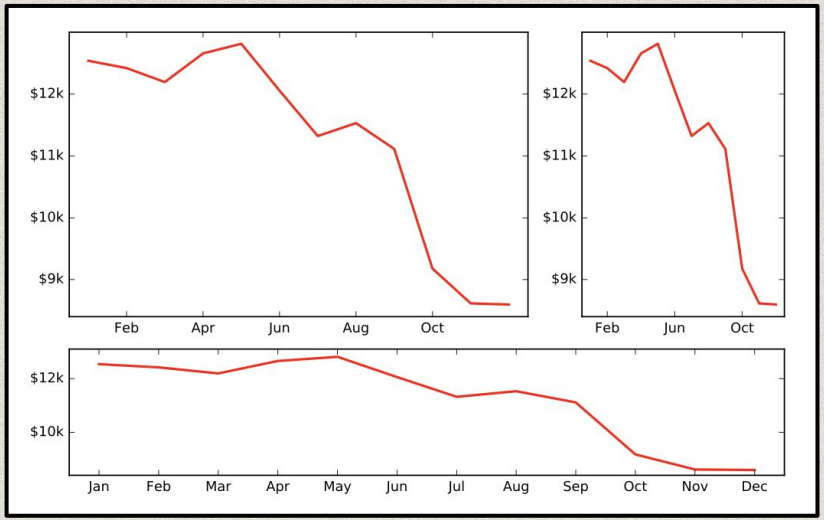


Edward Tufte's principles

Best practices

Minimize lie-factor

Omitting data or changing proportions in the visualisation can tell different stories.

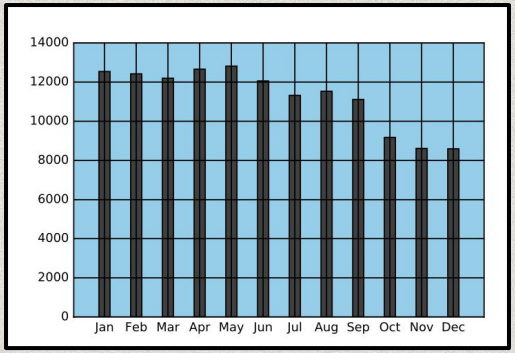


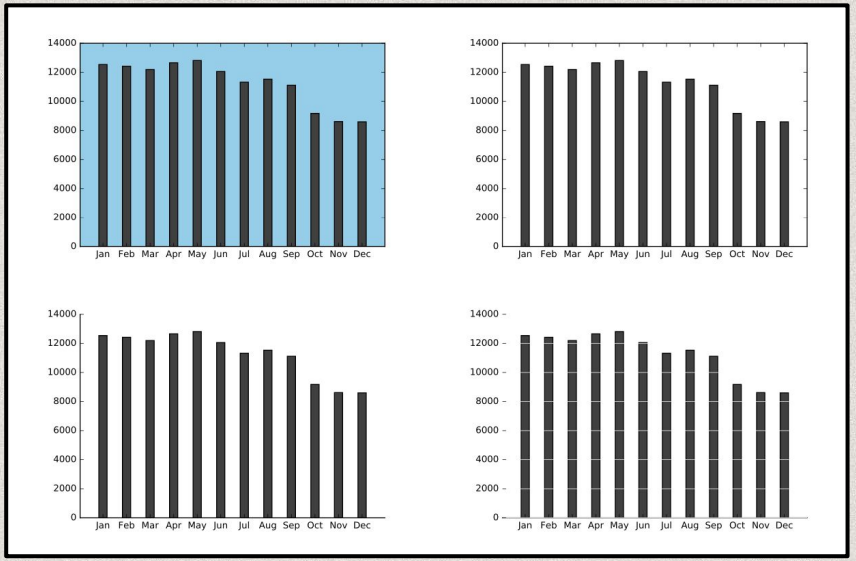
Edward Tufte’s principles

Best practices

Minimize chartjunk

Visual elements should not hide the data.



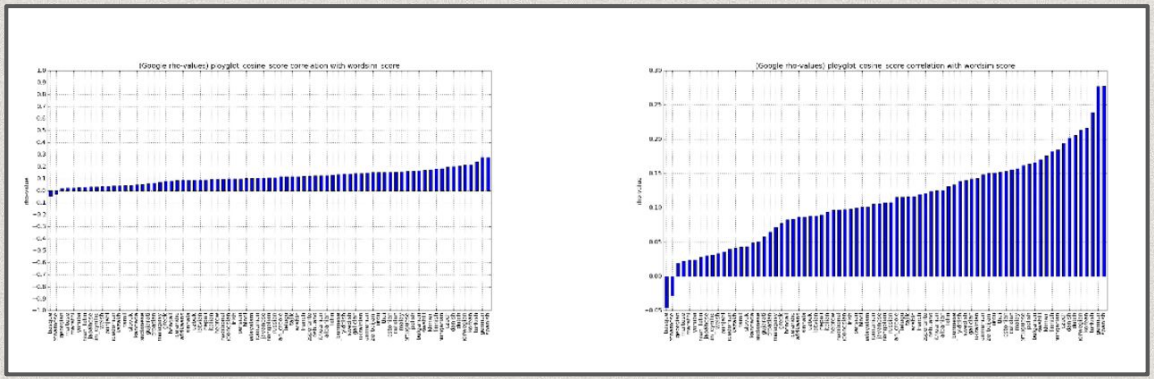


Edward Tufte’s principles

Best practices

Scale and ratio

Data should be scaled so as to fill all the space allotted to it. Labels should be high resolution.

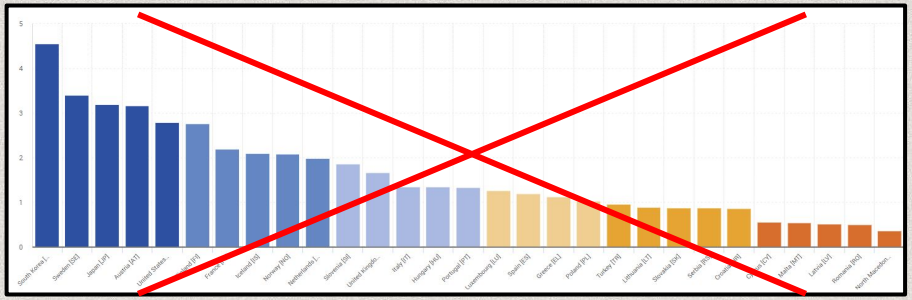


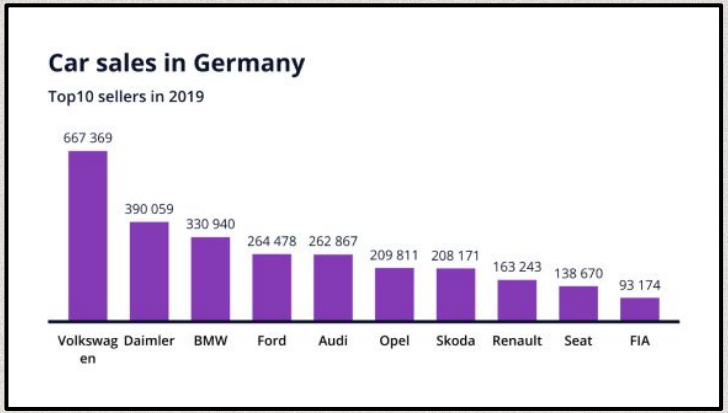
Edward Tufte’s principles

Best practices

Use of color

Use colors to distinguish values and represent **difference** - rather than representing linear, numerical scales.



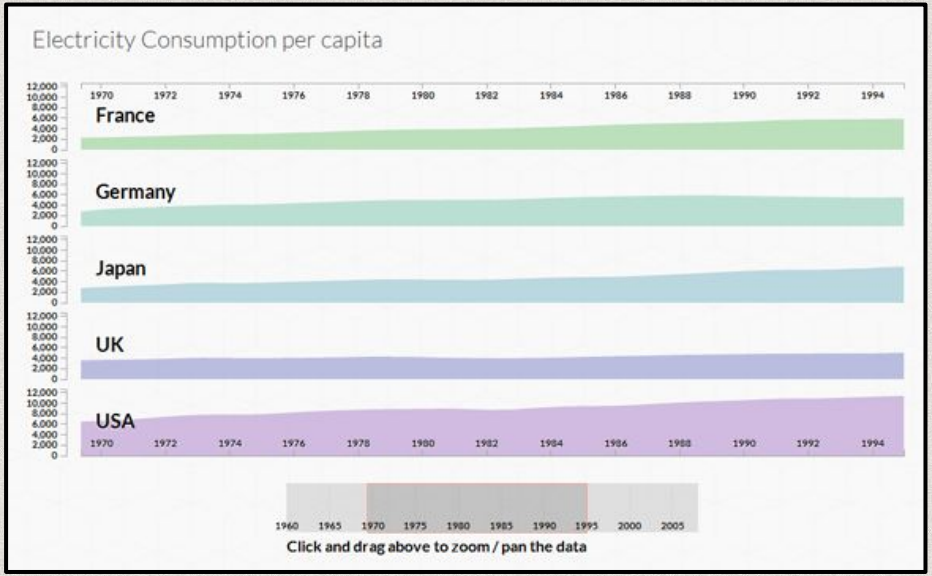


Edward Tufte's principles

Best practices

Repetition

Use multiple smaller charts to facilitate comparison (while losing the big picture)



04

Charts and graphs

An inspirational journey rather than a complete survey

Data Visualization types

04

Charts review

In the literature you will find many categorizations of charts (according to data types, number of variables, level of interactivity, task, affordance, etc.)

Tables

Graphs

Charts

Maps

Networks

Some catalogues

- Atlas
- Dataviz

Tables

Demonstrate results

If you want to scientifically **demonstrate** something, you will be asked to show a table.

It’s often the **starting point** to realize more sophisticated visualizations.

0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

0.103	0.176	0.387	0.300	0.379	0.276	0.179	0.321	0.192	0.250
0.333	0.384	0.564	0.587	0.857	1.064	0.698	0.621	0.232	0.316
0.421	0.309	0.654	0.729	0.228	0.529	0.832	0.935	0.452	0.426
0.266	0.750	1.056	0.936	0.911	0.820	0.723	1.201	0.935	0.819
0.225	0.326	0.643	0.337	0.721	0.837	0.682	0.987	0.984	0.849
0.187	0.586	0.529	0.340	0.829	0.835	0.873	0.945	1.103	0.710
0.153	0.485	0.560	0.428	0.628	0.335	0.956	0.879	0.699	0.424

Tables

04

Charts review

Precise comparison

- Good for **multivariate** variables
- Rows should invite **comparison** (e.g. sort table in ascending order by one column value)
- The order of columns represents their **importance**

Frequency (Hertz)	Scientific Pitch	Helmholtz/ German	Octave Name	Pipe Length
8.176	C-1	CCCC		64'
16.352	C0	CCC	sub-contra	32'
32.703	C1	CC	contra	16'
65.406	C2	C	great	8'
130.81	C3	c	small	4'
261.63	C4	c'	1-line	2'
523.25	C5	c''	2-line	1'
1046.5	C6	c'''	3-line	1/2'
2093.0	C7	c''''	4-line	
4186.0	C8	c'''''	5-line	
8372.0	C9	c''''''		

Tips: change background color, font size, alignment of values, add icons (e.g. ✖) to emphasise content.

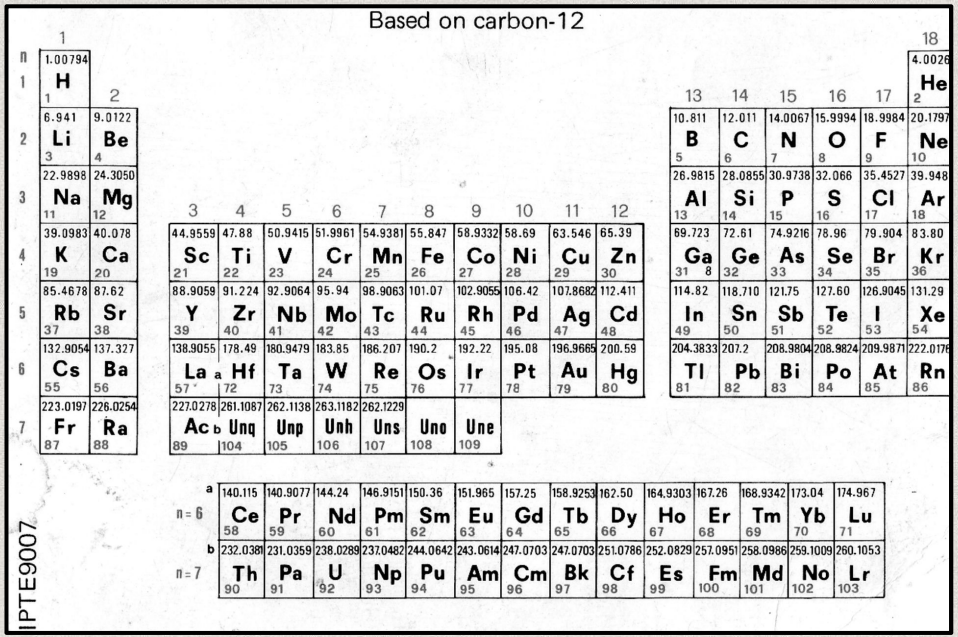
Tables

04

Charts review

Mendeleyev's table

- **Rows** (called periods) identify the level of energy
- **Columns** (called groups) gather elements with similar properties (power law)
- The **horizontal order** of elements represent the incremental atomic number (n. of protons in the nucleus)



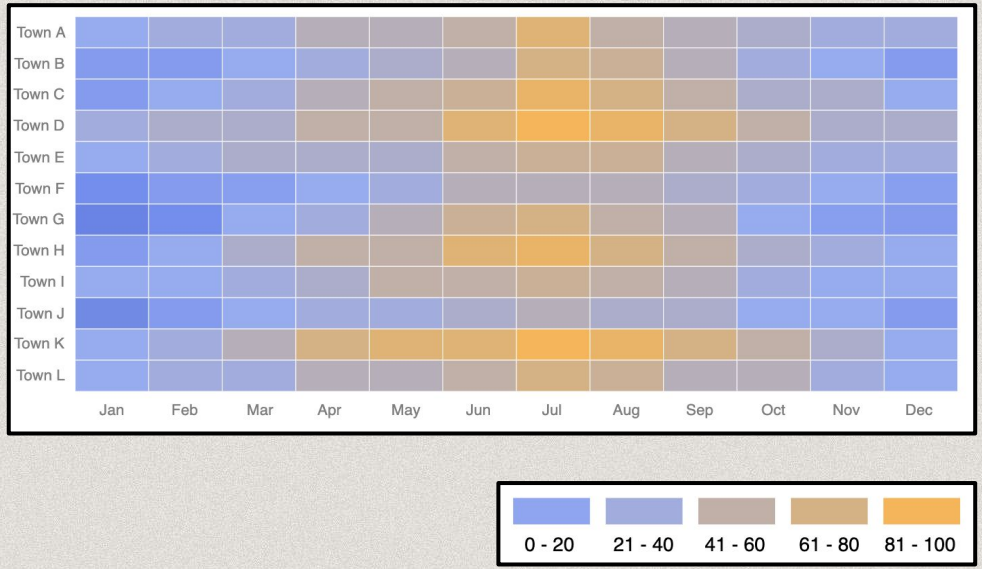
Tables

04

Charts review

Heatmaps

- Good for **multivariate** (3) variables
- Numeric values are mapped to colors
- It shows the **variance** (gradation) across values and can highlight patterns (e.g. correlation between month and temperature)
- Not good for precise comparison

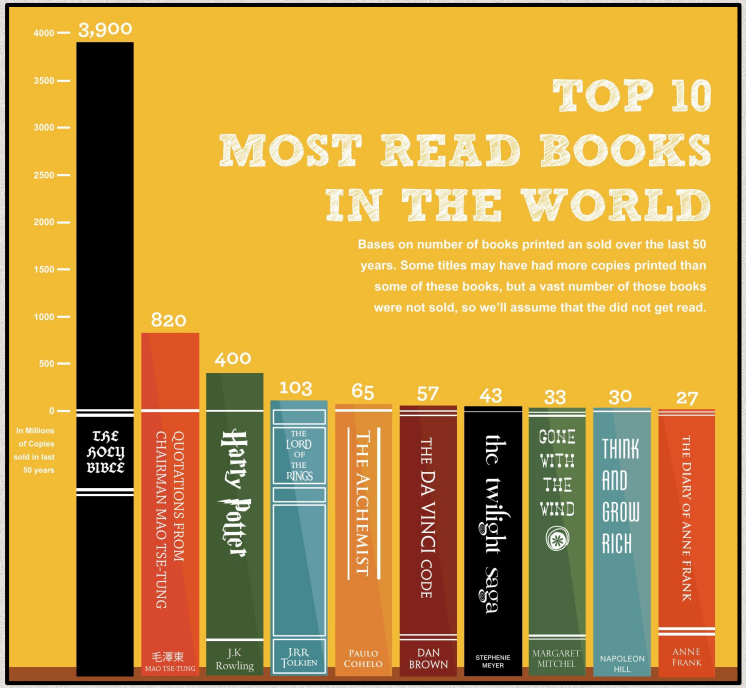


Graphs

Charts review

Bar graph

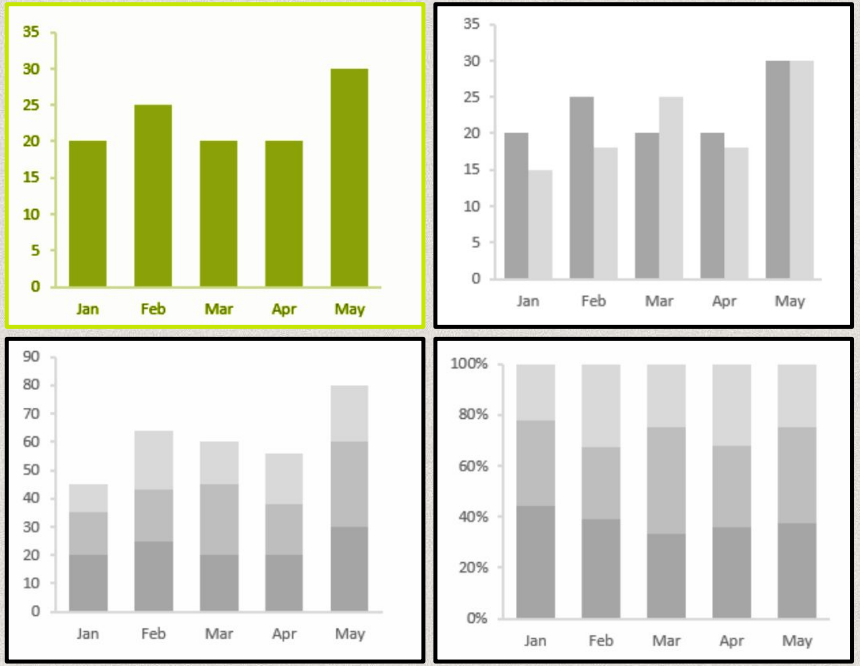
- Good for **bivariate** (2) analysis, often including
  - 1 categorical or ordinal variable
  - 1 numeric variable
- Good for **comparison** between (distribution of) categories



Graphs

Bar graph

In column graphs the categorical/ordinal variable is **independent** (x axis), while the numeric one is **dependent** (y axis)



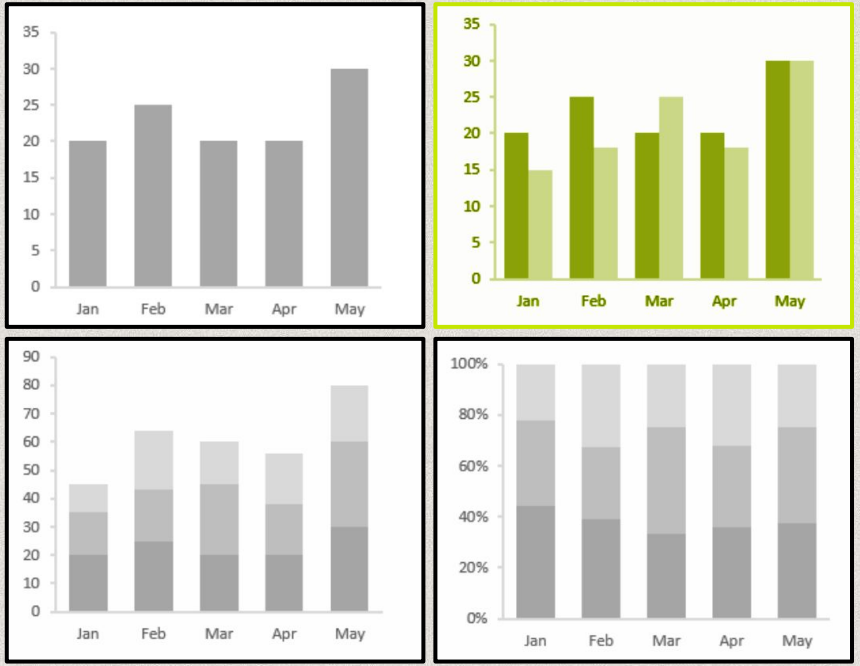
Graphs

Charts review

Multi-series bar graph

When we need to compare sub-categories falling under a macro-category, we have **multi-series** or grouped bar graphs.

Sub-categories are differentiated by colors.

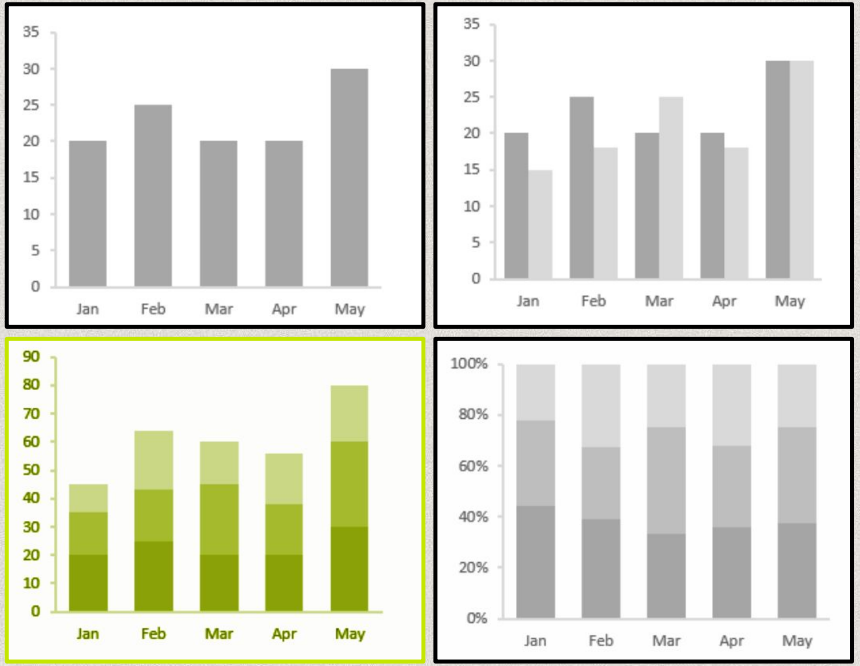


Graphs

Stacked bar graph

Similarly to grouped bar graphs, **stacked bar graphs** allow to compare sub-categories. These are placed on top of each other rather than next to each other.

Sub-categories are differentiated by colors.

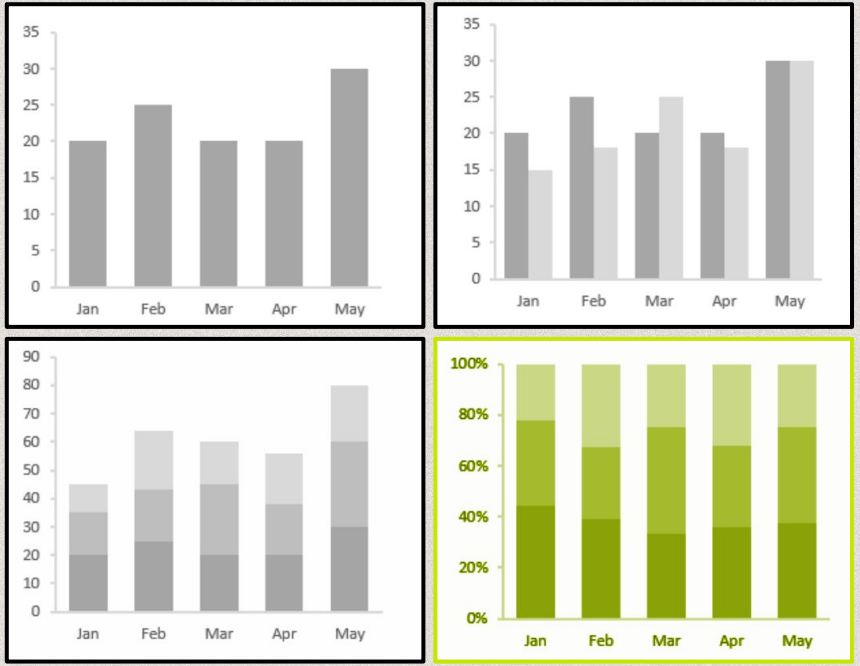


Graphs

100% Stacked bar graph

100% stacked bar graphs or normalised stacked bar graphs, allow to compare the contribution of sub-categories to the category they belong to (proportion).

Sub-categories are differentiated by colors.

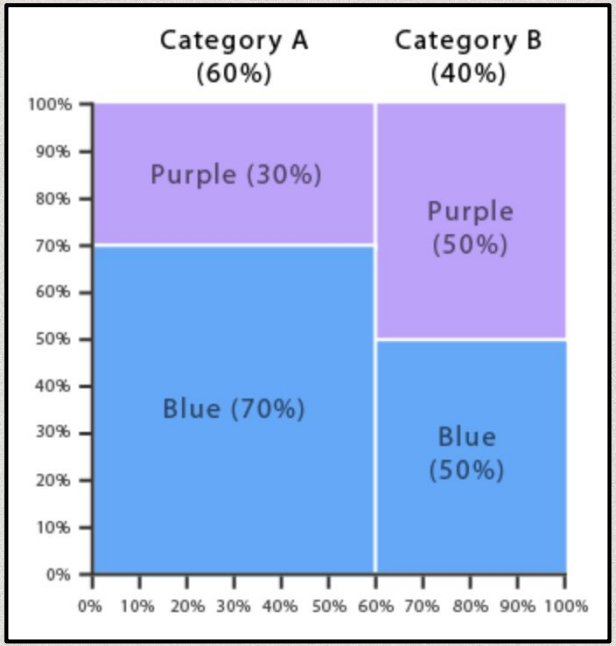


Graphs

Charts review

Marimekko

A special 100% stacked bar graph, wherein not only the height but also the width of columns grows according to the proportion of sub-categories wrt the macro-category.



Graphs

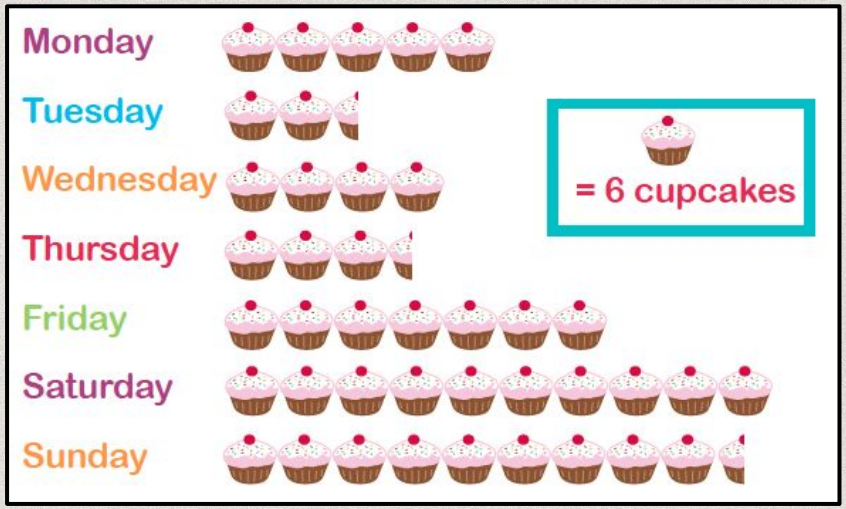
04

Charts review

Pictograms

Pictograms are special bar graphs where columns are replaced with **icons**.

It is a more engaging way to invite **comparison**.



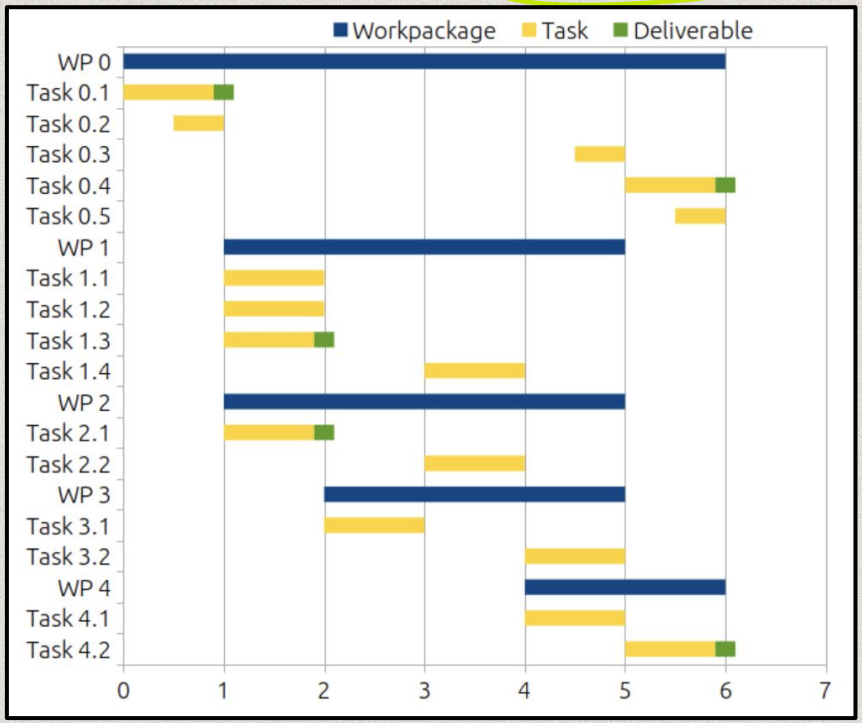
Graphs

Charts review

Gantt chart

Halfway between a table and a bar chart (wherein the origin of axes may change for every column), represents variables over a time span (the ordinal value on the x axis).

It allows to see **dependencies** between variables.

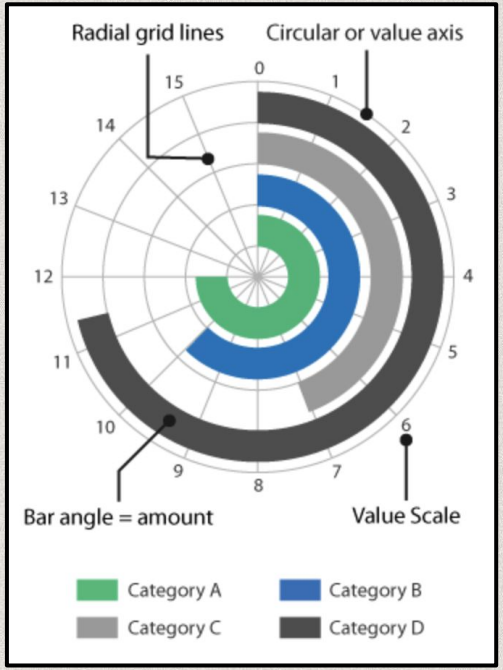


Graphs

Radial bar chart

A bar chart where the cartesian system is replaced by a **polar system**.

Unfortunately, they are easy to be **misread**. Outer bars generally tend to seem bigger than inner bars, while it does not imply that is true (see the example).

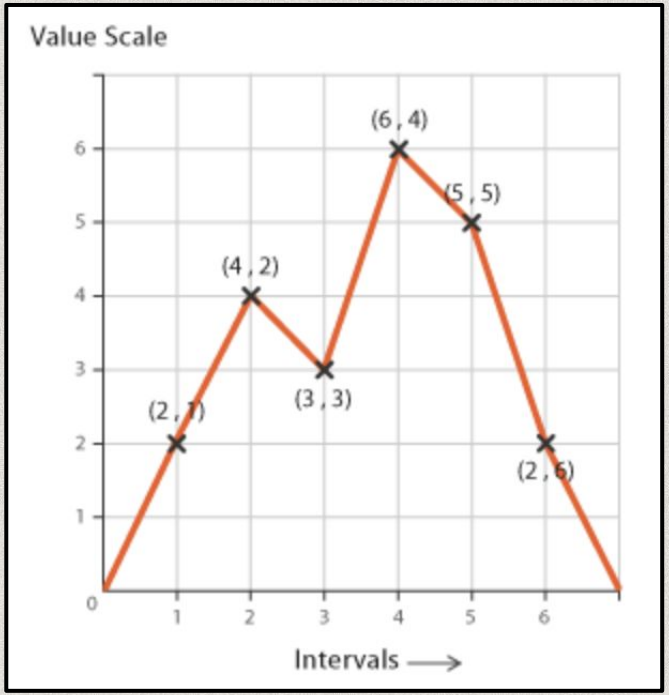


Graphs

Charts review

Line graphs

- Good for **bivariate** analysis
- Like in bar graphs, one variable can be categorical/ordinal
- Variables are always **dependent**
- For each category only one point is shown. The line can be drawn or not
- Good to show **time-series** and highlight **trends**.



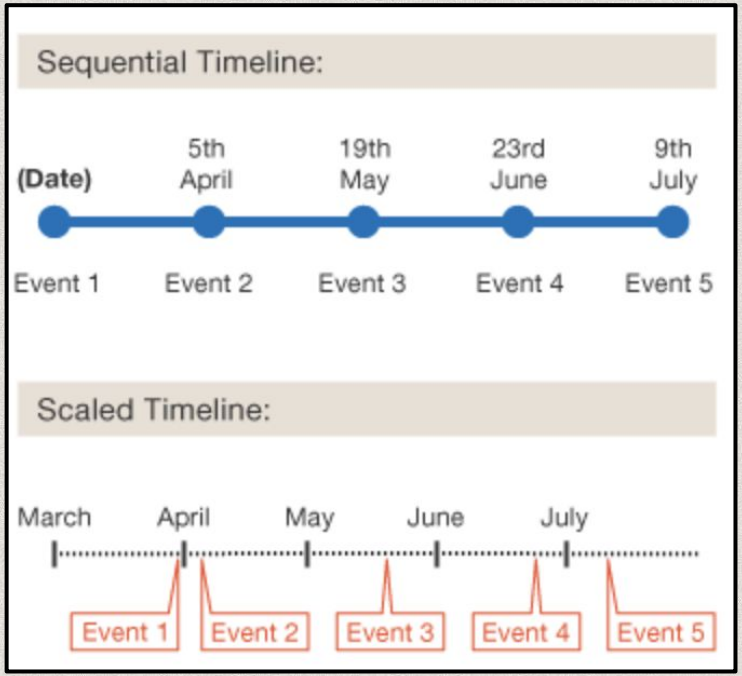
Graphs

Timeline

Data points can be displayed on a vertical/horizontal line.

The distance between points can be according to a **scale** or in **sequence** (no scale).

In a scale-based timeline it's possible to see the **clusters** of data points.



Graphs

Area charts

A special line chart where colors highlight the **dependency** between 2 variables.

It both shows the **trend** and the **size** of a phenomenon.

Values are scaled (e.g. percentage).

Graphs

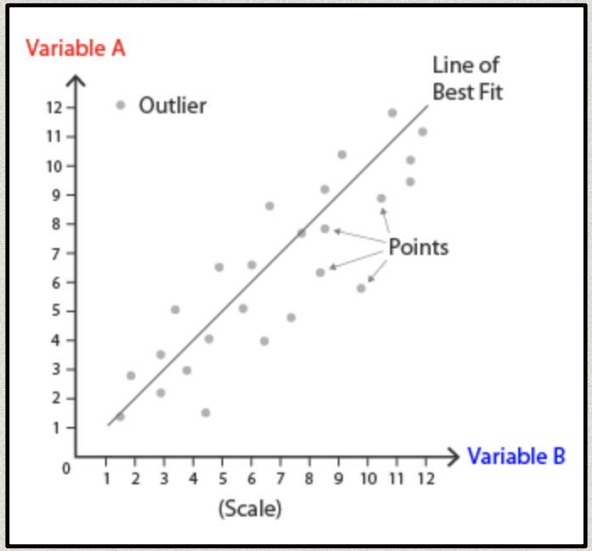
Charts review

Scatter plots

- Allow **bivariate** analysis between numeric values.
- Show the **correlation** between two numeric variables and the **outliers**.

Correlation is a measure that determines the degree to which the movement of two different variables is associated.

If a correlation exists, a line (called **best fit**) can be drawn.



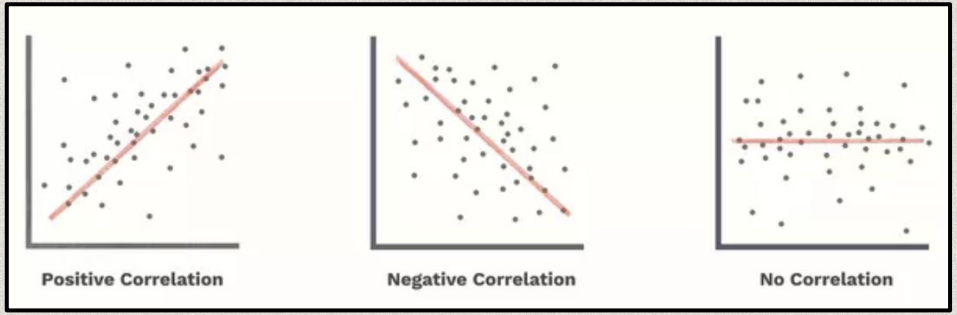
Graphs

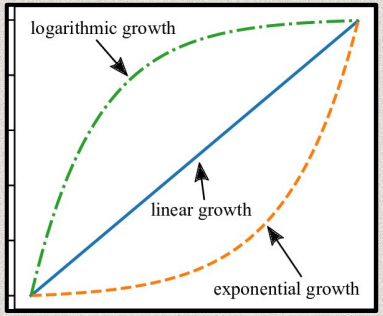
04

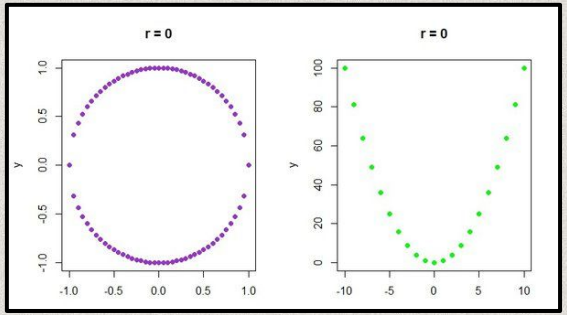
Charts review

Scatter plots

- Patterns (highlighted by a **best fit** line) include:
  - positive (values increase together),
  - negative (one value decreases as the other increases),
  - null (no correlation),
  - linear
  - exponential
  - U-shaped.







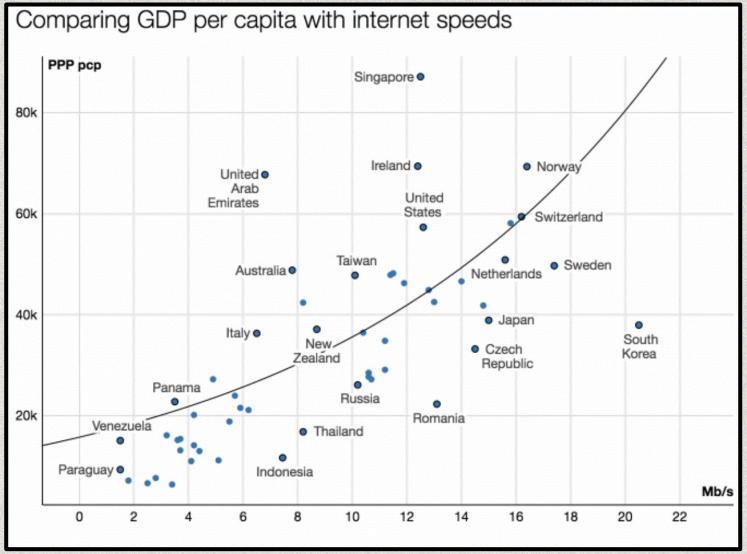
Graphs

04

Charts review

Positive correlation examples

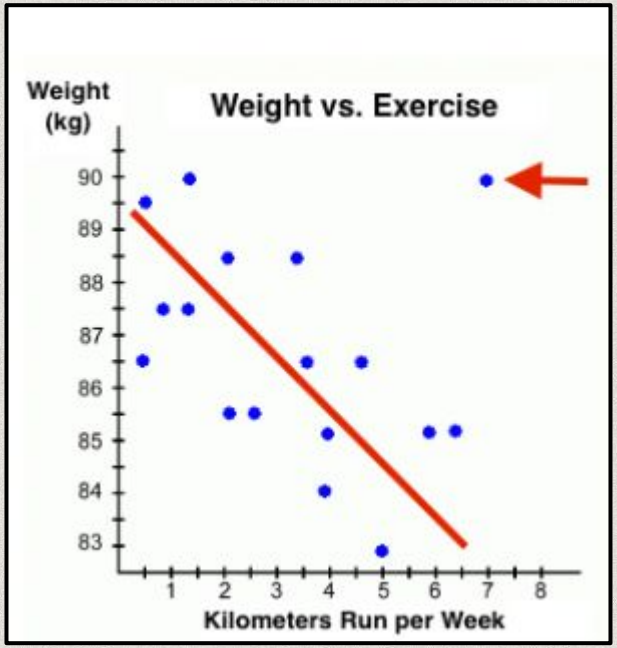
The more the income of a country grows (y axis), the more the speed of wifi grows (x axis) - and vice versa.



Graphs

Negative correlation examples

The more a person exercise (x) the less a person weights (y).



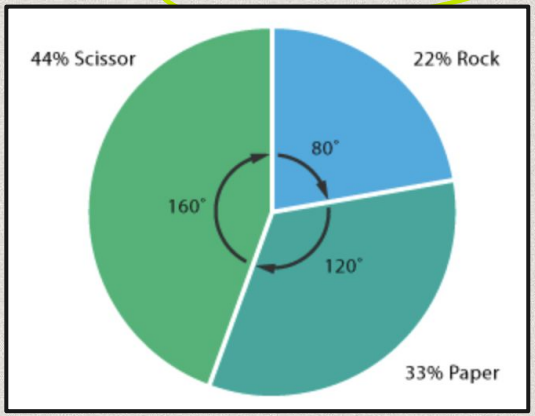
Charts

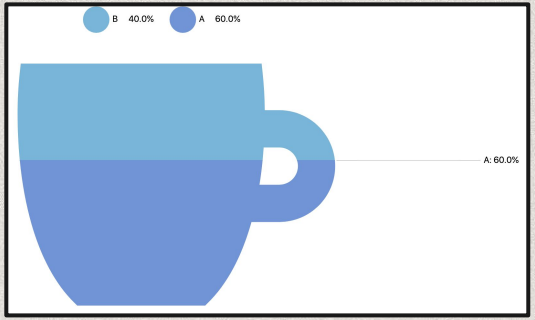
Charts review

Pie chart

- Shows the **proportion** of parts (categorical values) of a whole.
- Good for **univariate** analysis with **dependent** values (if one changes also others change).
- **Angles** do not allow a precise comparison

Circles can be replaced with images.





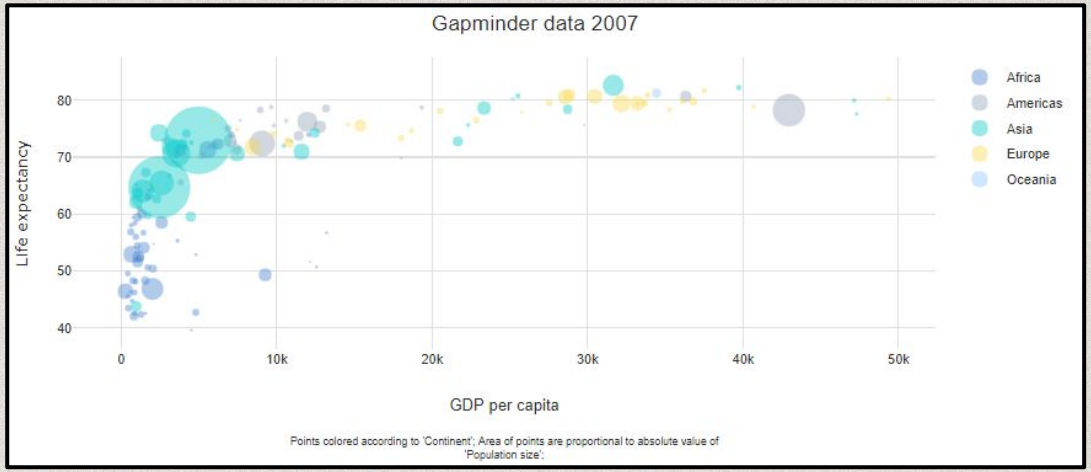
Charts

04

Charts review

Bubble chart

- Good to compare a great amount of **multivariate** data
- Position and proportion can reveal **correlation** (e.g. here is logarithmic)



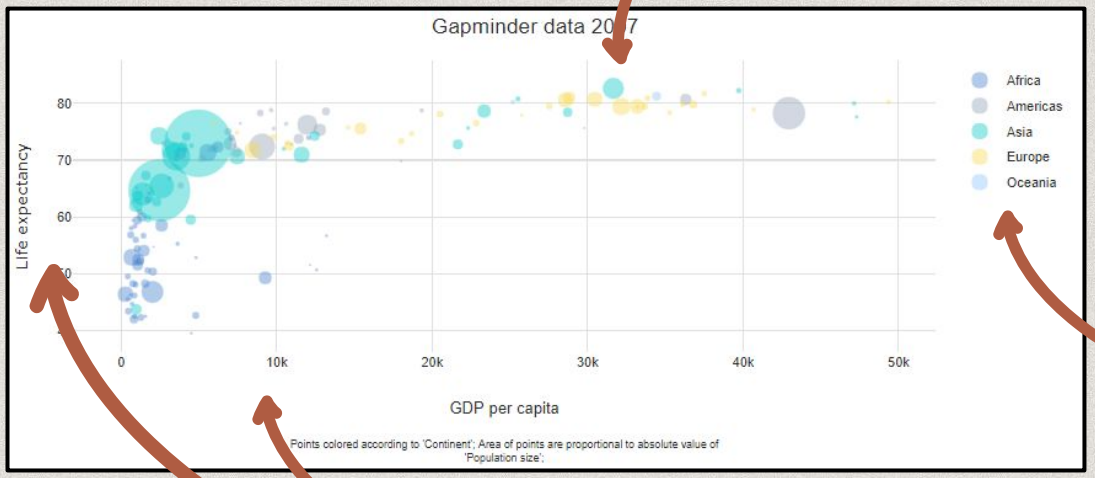
Charts

04

Charts review

Bubble chart

The area is a numerical value



Points are labelled with a categorical value (country)

The color of points can represent another categorical value (continent)

X and Y are numerical values

Charts

Charts review

Wordle

- the size of words represents a **univariate** quantitative aspect of categories.
- The layout is based on a randomised greedy algorithm that positions words so that they fill empty spaces and words do not overlap.

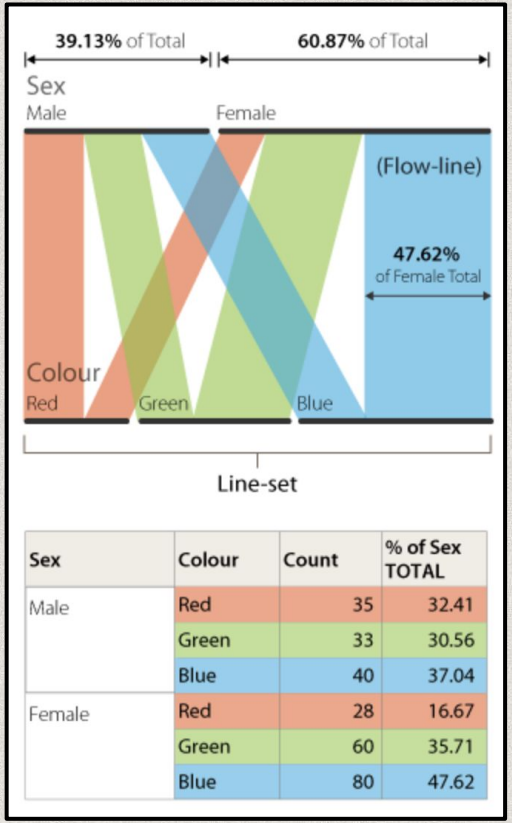
Tag clouds can be **misleading** (long words, letters with ascending lines seem bigger) and do not allow precise comparison.

Charts

Parallel sets

Allow to compare **proportions** and relations between **multivariate** categorical values.

Relations are shown as flows between pairs of values.



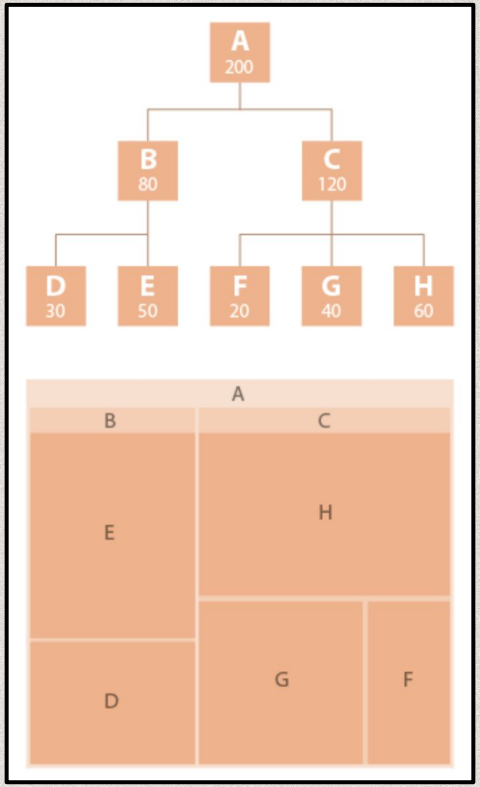
Sex	Colour	Count	% of Sex TOTAL
Male	Red	35	32.41
	Green	33	30.56
	Blue	40	37.04
Female	Red	28	16.67
	Green	60	35.71
	Blue	80	47.62

Maps

Tree map

Like pie charts, a tree map shows the **proportion** of parts of a whole.

It is a different way to display a **hierarchy**.

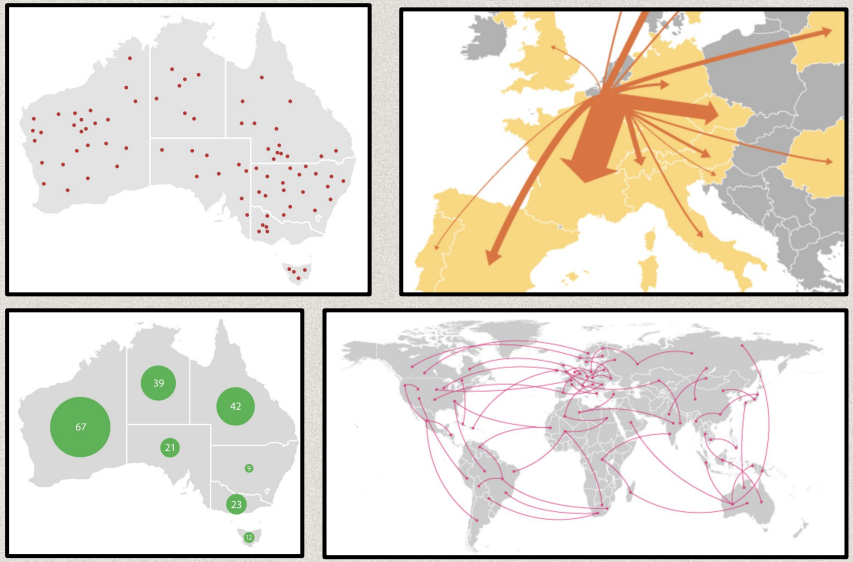


Maps

Geographical map

Data that can be **geo-localised** can be plotted on a map.

Maps can be combined with other visual strategies (e.g. flows, trajectories, dots, bubbles).



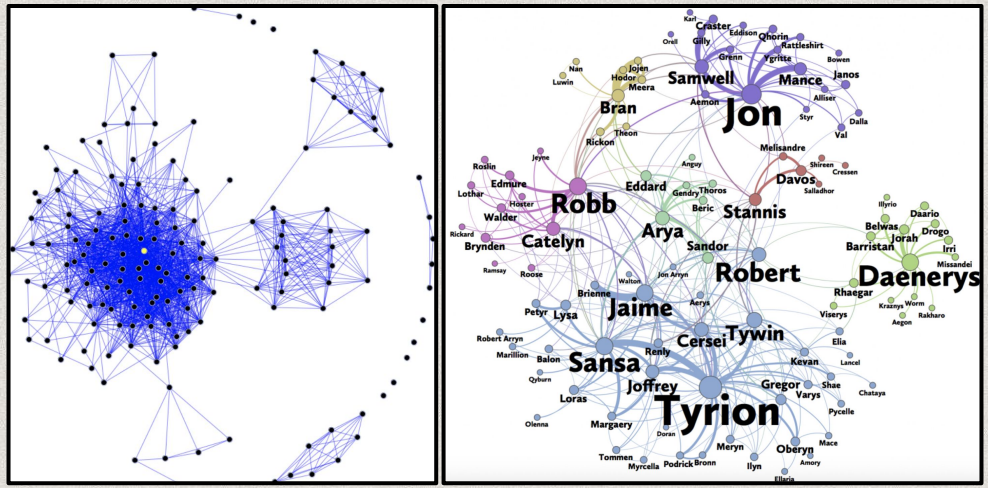
Networks

Graph networks

Categorical data are mapped to **nodes** and relations (co-occurrence) become (weighted) **arcs** between nodes.

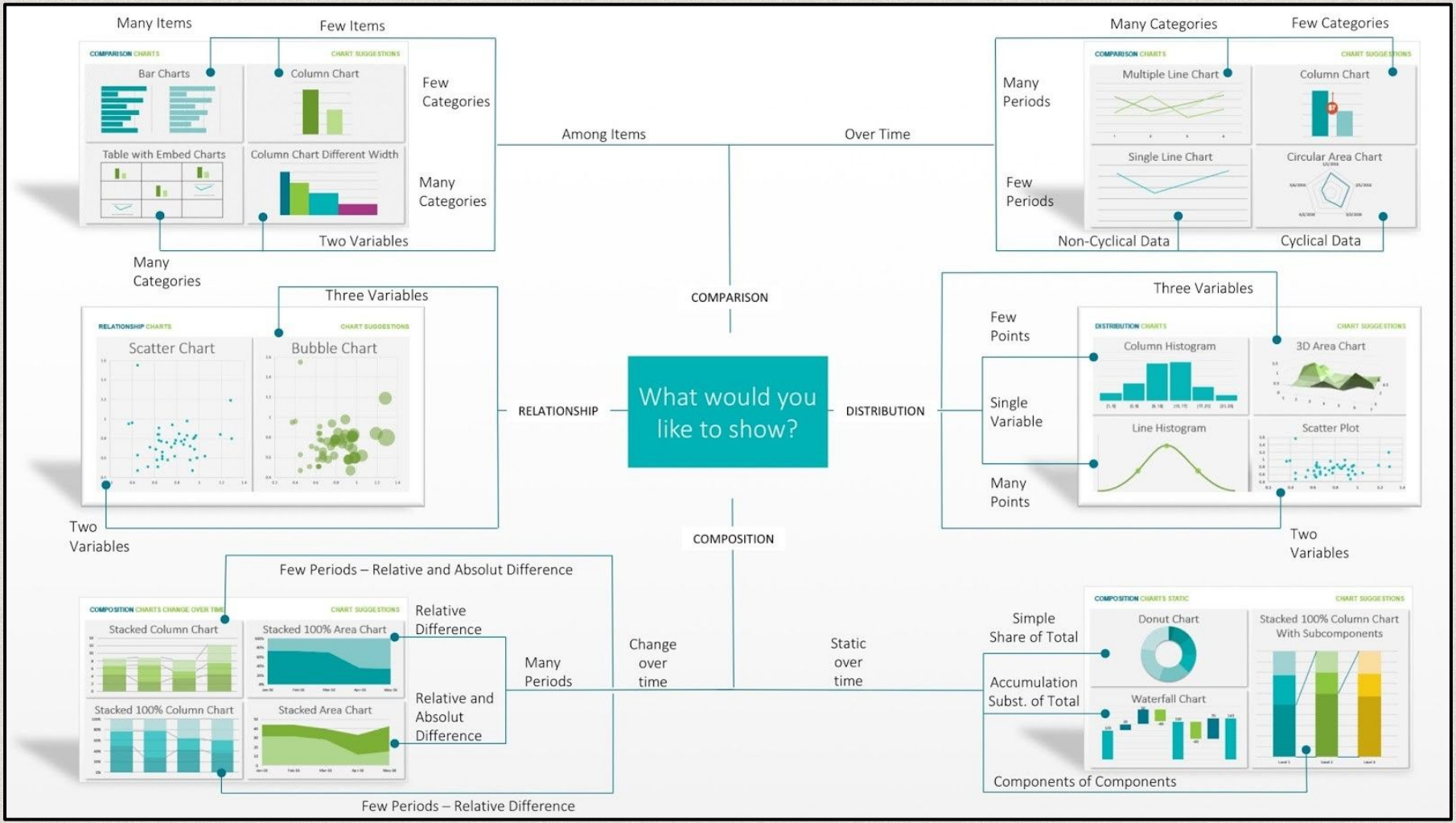
The location (closeness) of clusters is relevant, showing **similarities** between groups of data.

They are unreadable :(



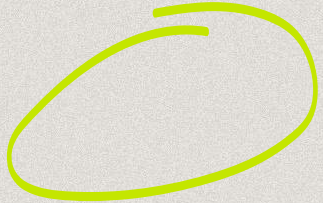
04

How to decide?



https://www.techprevue.com/decision-tree-perfect-visualisation-data/

Exercise



!!!

EDA

Review

Review these slides at home

Answer

Fill in the questionnaire

TODO

Come prepared next week!

Bring your laptop if possible

Install Python3 and a IDE (e.g. PyCharm)

Know to run a py script via shell (read here)

Know how to install a Python library with pip (or install RDFlib)

Thanks!

Do you have any questions?

marilena.daquino2@unibo.it

https://github.com/marilenadaquino/information_visualization