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

Design issues

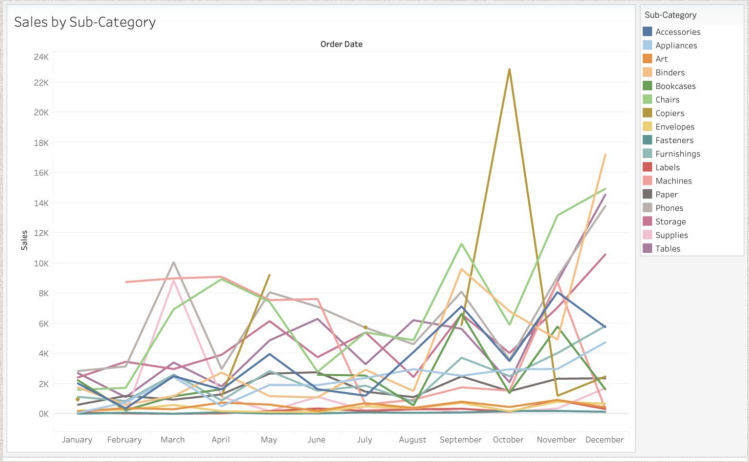
Terminology, expectations and best practices do design effective visualisations

Types of analysis

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

The visual summary should be able to:

- **explore** a dataset (give an overview)
- **discover** knowledge (answer a question)



An example of graph that identifies and summarises data, but struggles to prioritise information.

01

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 it good for? which types of analyses does it enable?).

In presentations, EDA is used as introduction and summary.

Knowledge discovery

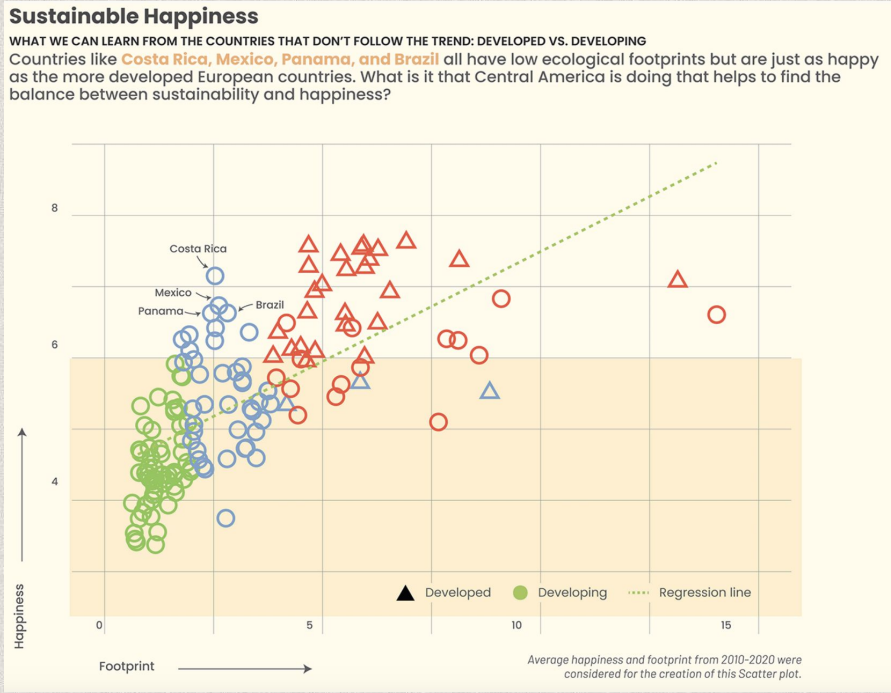
Objectives

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

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

The validation consists of the **visual evidence** of a trend, a behaviour, an outlier, or a relationship (i.e. a pattern).

source



Results of data analysis

Results of an analysis must be:

- **Valid** (representative): hold true when new data are added (if completeness cannot be ensured).
- **Novel**: are non-obvious.
- **Actionable**: can be used in practical tasks (e.g. decision-making).
- **Understandable**: are interpretable and meaningful to humans.

Expectations on novelty

Outcomes of data analysis/discovery techniques fall into two dichotomies: **expected/unexpected, positive/negative**

Health and a good diet are positively correlated	Expected positive	Unexpected negative
Health and junk food are negatively correlated	Expected negative	Unexpected positive

Gender affects the chances to get a loan

Playing with less toys leads children to play longer

Expectations on novelty

Unexpected negative results are the most insightful, **actionable** patterns to support decision-making (something you can act upon to change the future).

Expected positive	Unexpected negative
Expected negative	Unexpected positive

Negative results

A terminological mismatch.

Negative (or null) results refer to those insights that contradict the initial hypothesis or that do not show enough statistical significance to confirm/deny a hypothesis.

Possible reasons:

- The original hypothesis was incorrect
- The findings cannot be replicated
- Technical problems

While they have often been discredited by the scientific community (i.e. it’s difficult to publish them, especially if data don't support an interesting alternative hypothesis) they provide interesting insights that could influence future analyses.

False/True Positive/Negative results

	Positive	**Negative**
False	A condition is demonstrated to hold while it actually does not.	A condition is demonstrated not to hold while it actually does.
True	A condition is demonstrated to hold and it actually does.	A condition is demonstrated not to hold and it actually does not.

Examples

Negative results.

False negative.

False positive.

1. A research shows that during COVID, whether you were an avid video game player or not, this factor did not affect your mental health overall.
2. Another research, based on a **different sample**, shows that video games did affect your mental health. Study n.1 was a false negative.
3. Another research, that has a **stronger methodological approach** in selecting the study population, shows that video games do not affect you. Study n.2 is a false positive.

Edward Tufte’s principles

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

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

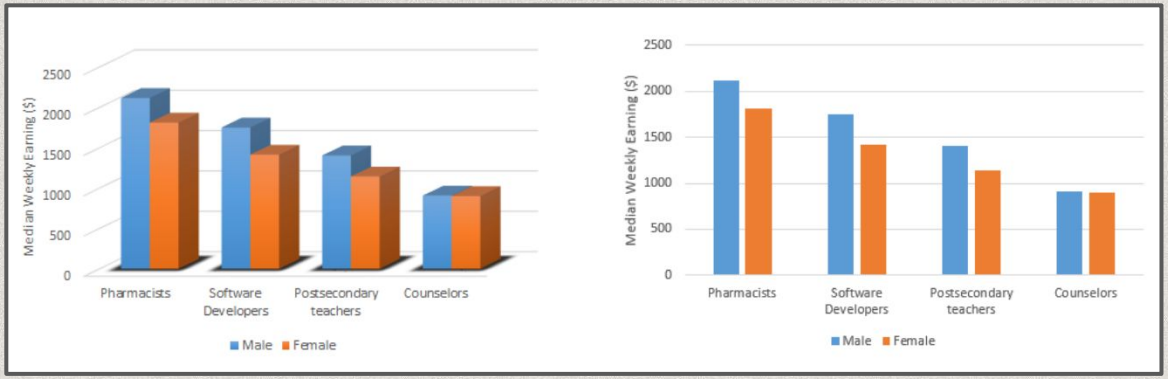
Skiena S. Data Science Design Manual. Springer. 2017

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.

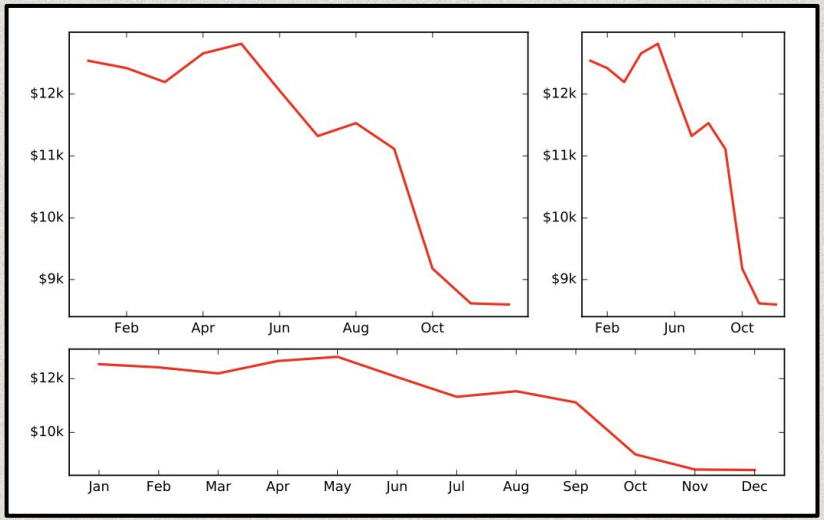


Edward Tufte's principles

Best practices

Minimize lie-factor

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



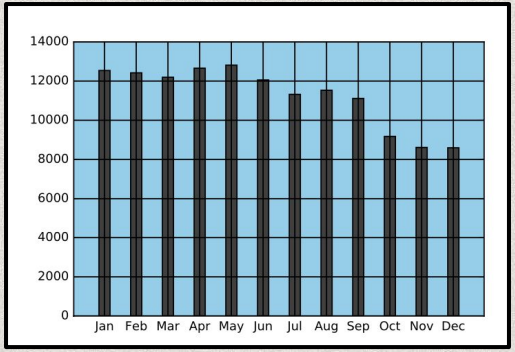
01

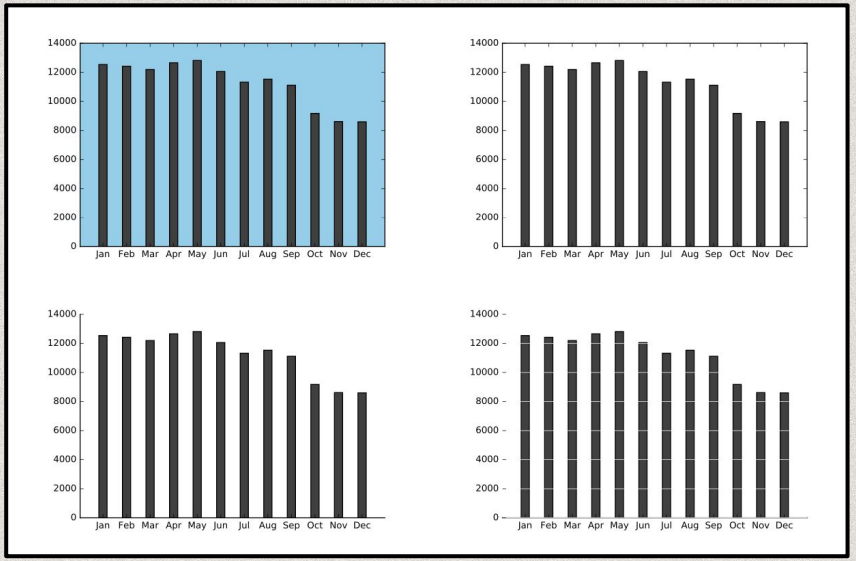
Edward Tufte’s principles

Best practices

Minimize chartjunk

Visual elements should not hide the data.



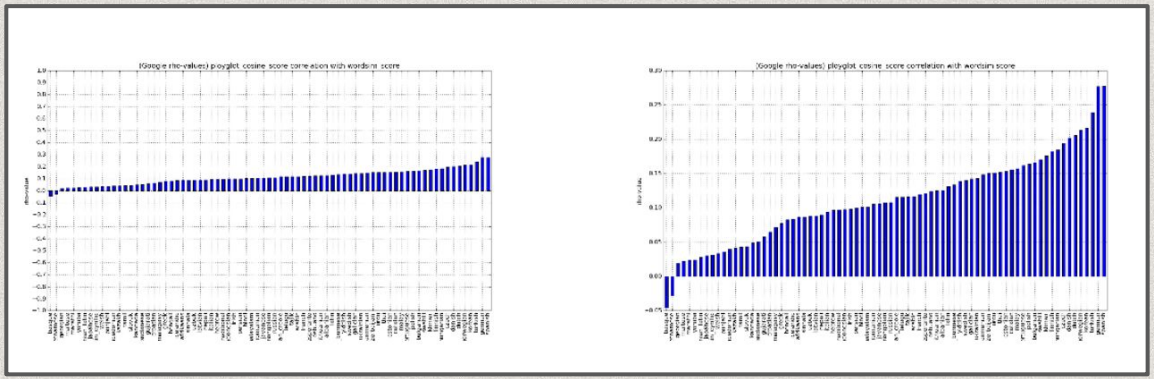


Edward Tufte’s principles

Best practices

Scale and ratio

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

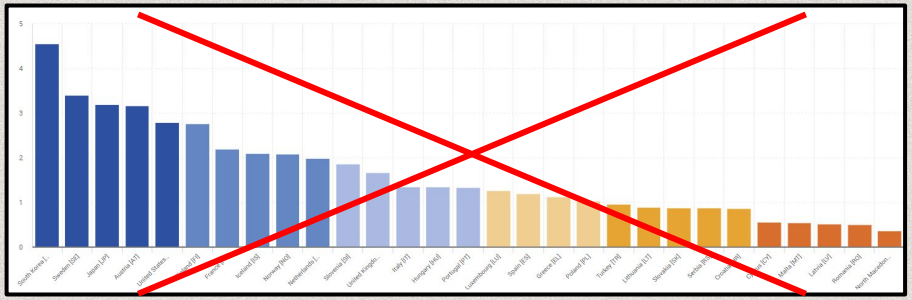


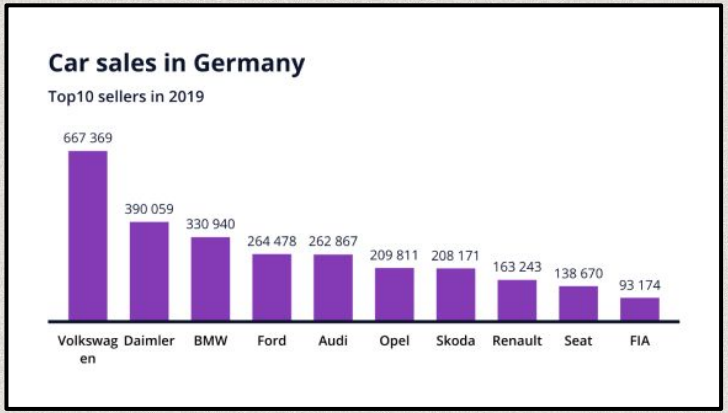
Edward Tufte’s principles

Best practices

Use of color

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



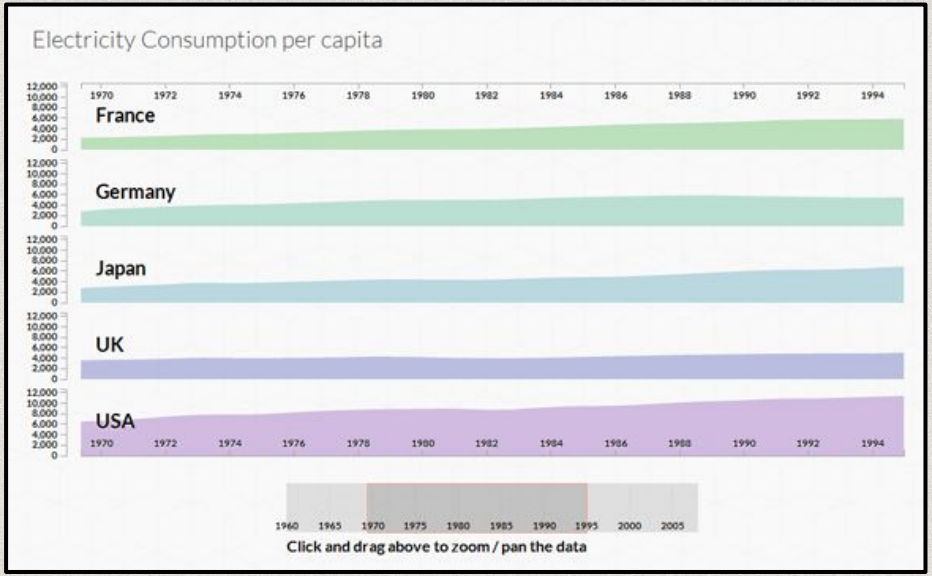


Edward Tufte's principles

Best practices

Repetition

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



02

Charts and graphs

An inspirational journey rather than a complete survey

Data Visualization types

02

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

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.

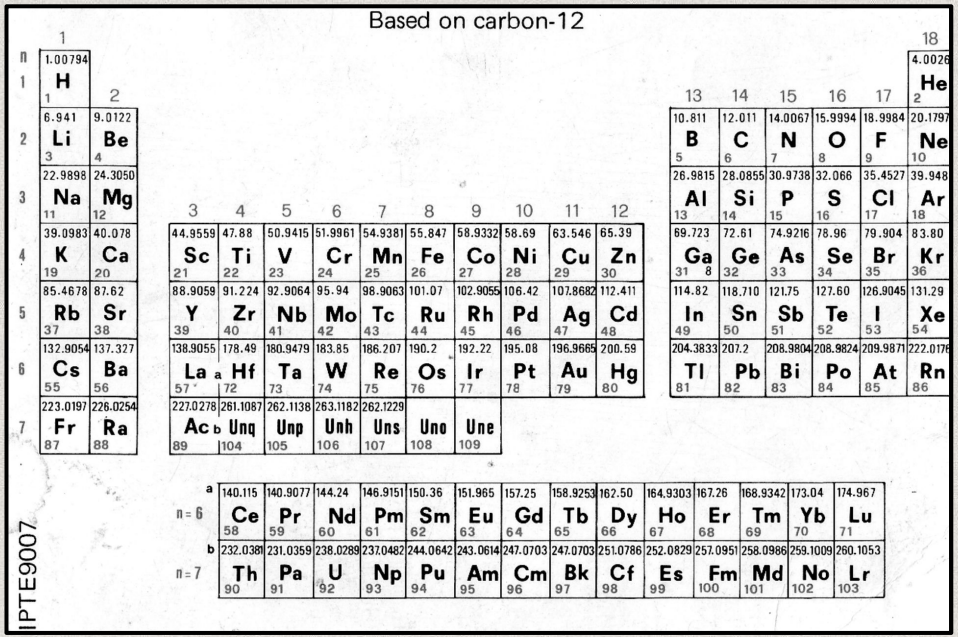
Tables

02

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)

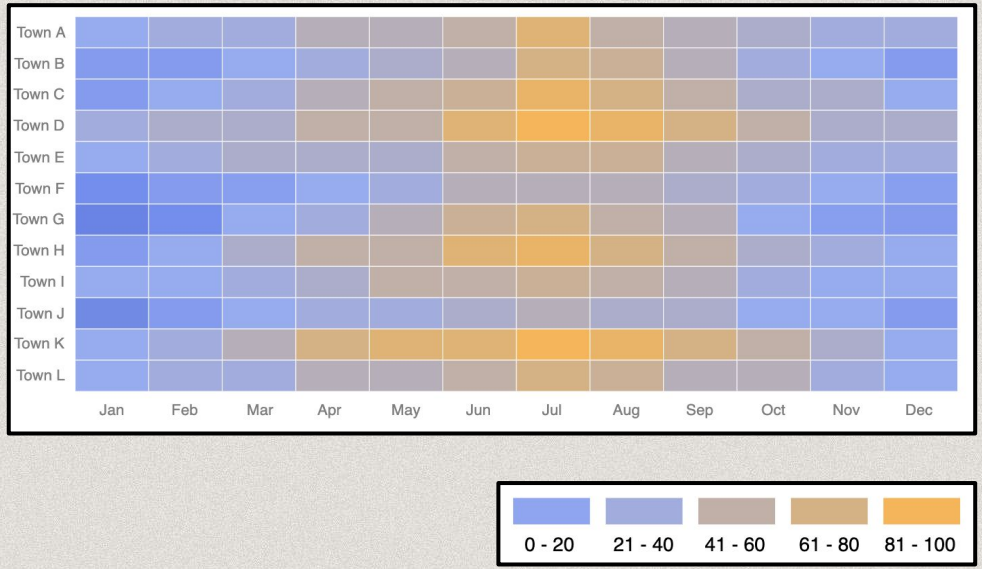


Tables

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



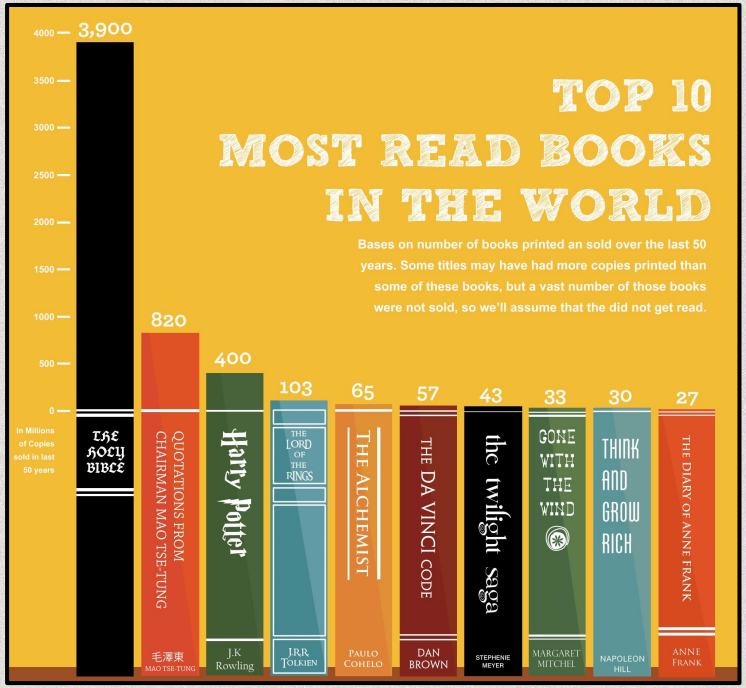
Graphs

02

Charts review

Bar graph

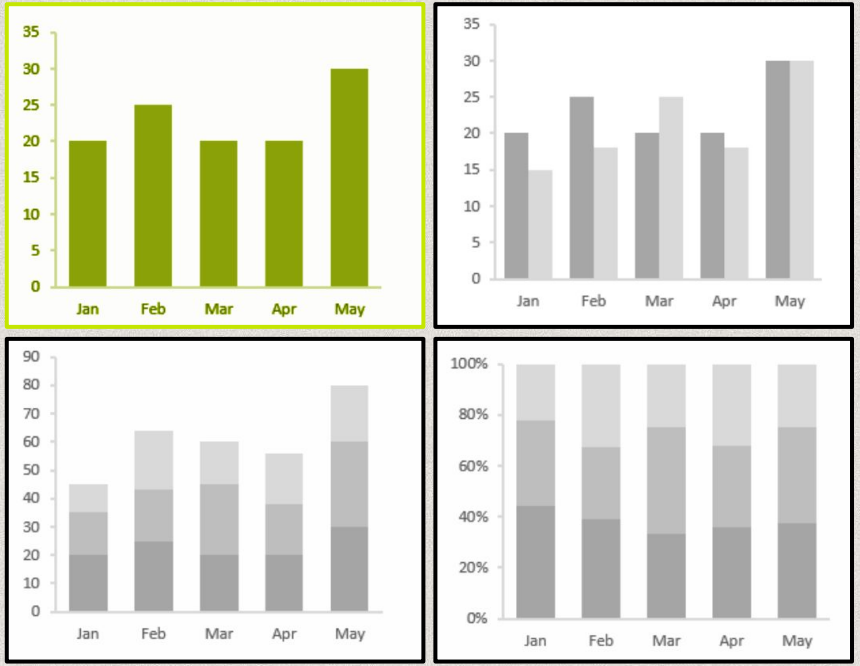
- Good for **bivariate** (2) analysis, often including
  - 1 categorical or ordinal variable
  - 1 numeric variable
- Good for **comparison** between (distribution of) categories



Graphs

Bar graph

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



Graphs

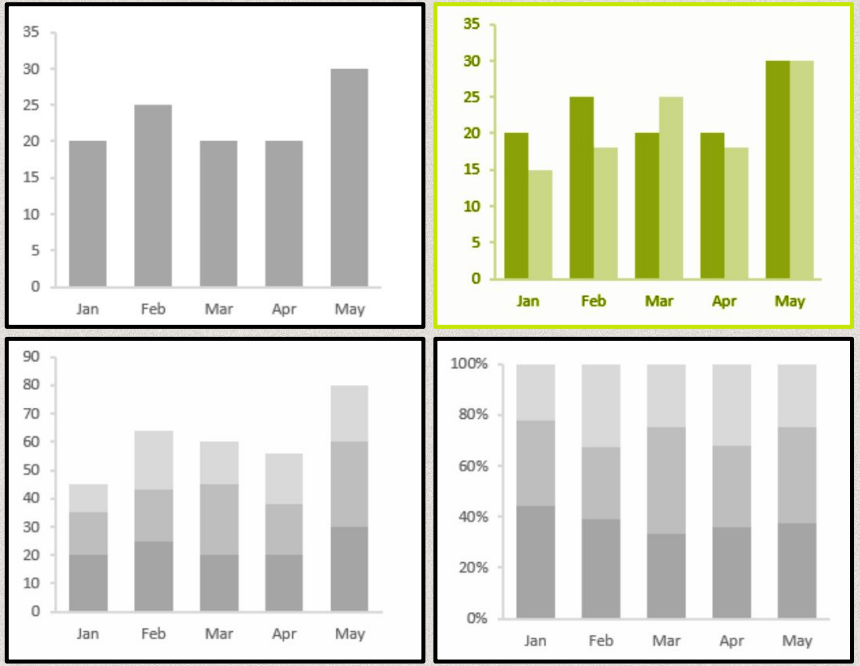
02

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.

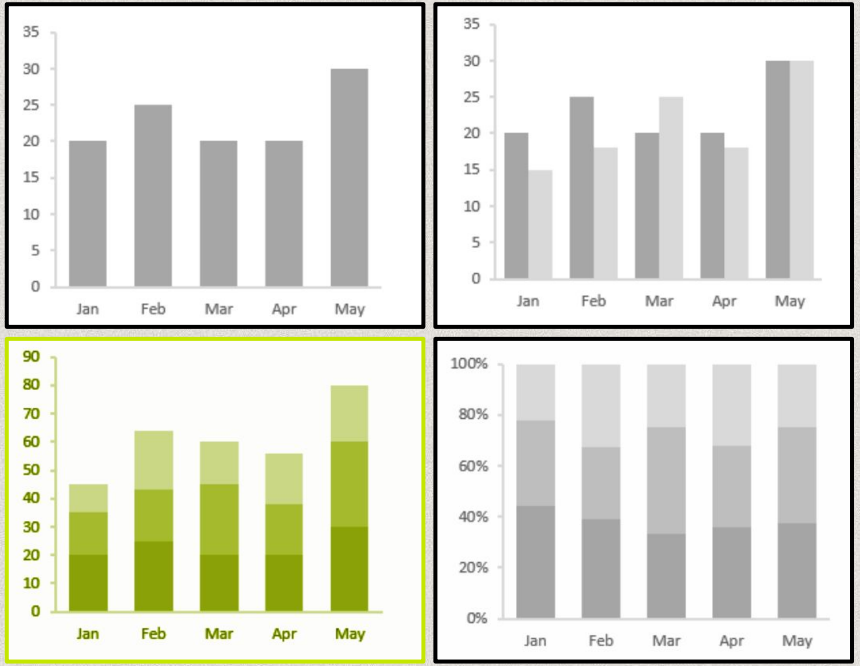


Graphs

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

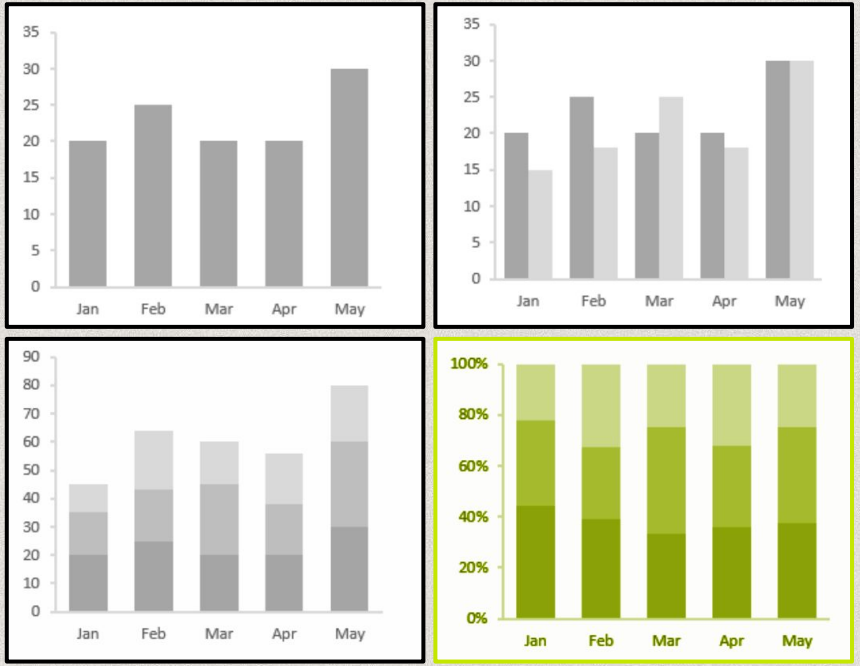
02

Charts review

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.



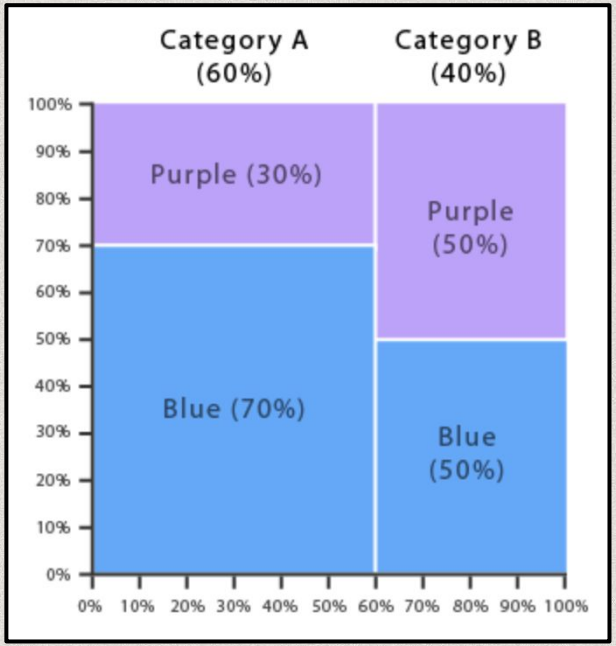
Graphs

02

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

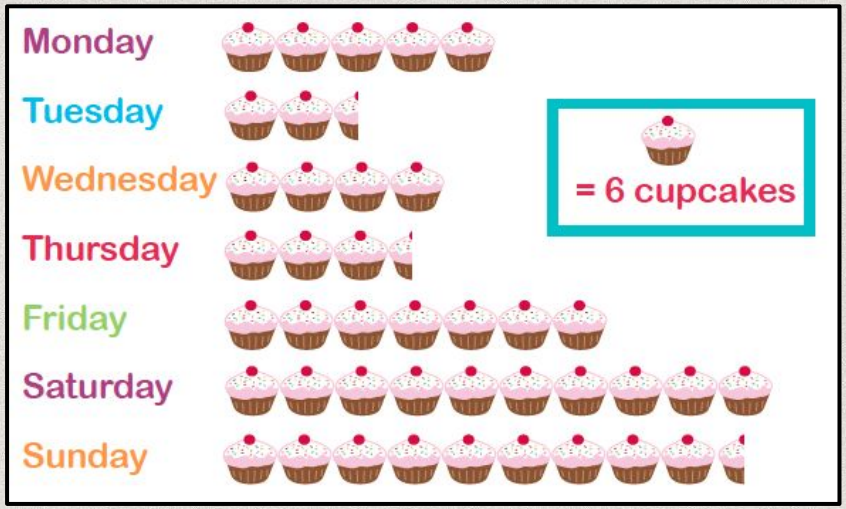
02

Charts review

Pictograms

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

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



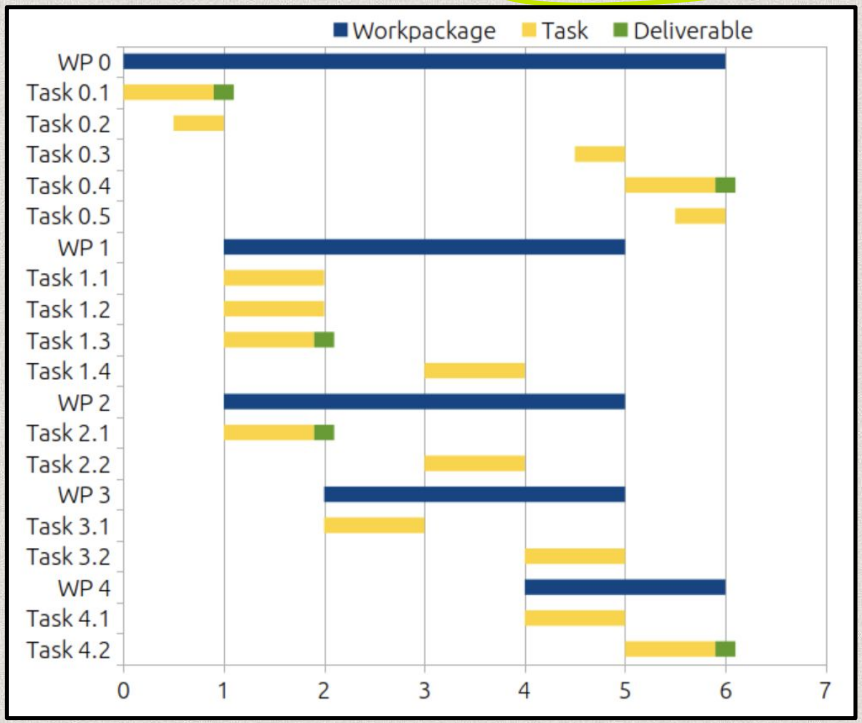
Graphs

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.

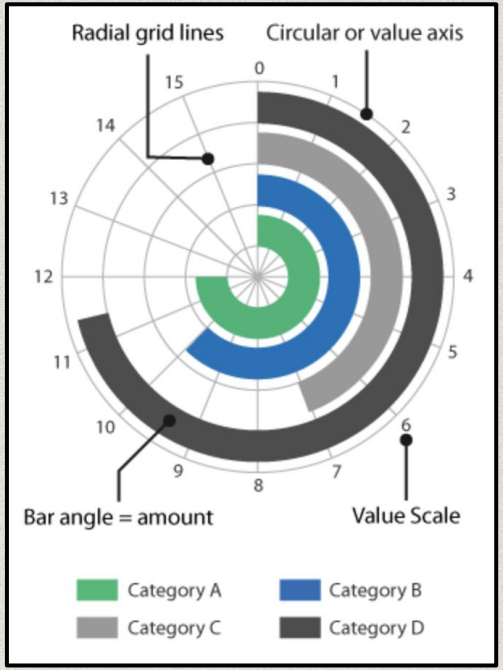


Graphs

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).



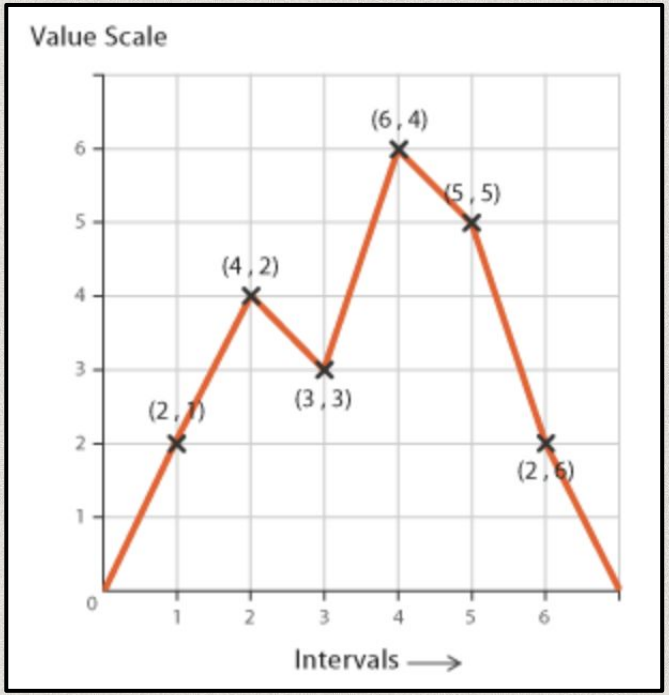
Graphs

02

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

02

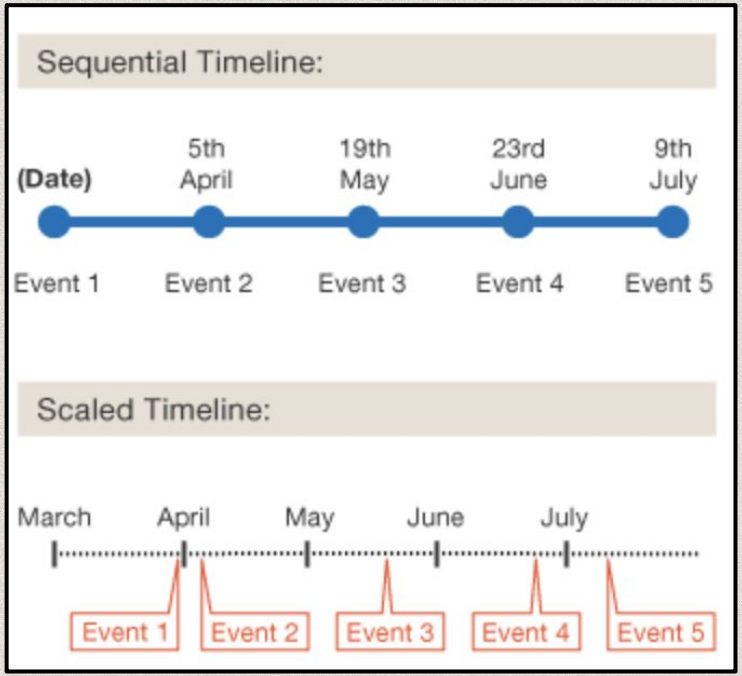
Charts review

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.



Data visualisation sense making

02

Test

Answer the questionnaire (20 minutes)

https://forms.gle/kchdc6fbP9CT82wx7

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

02

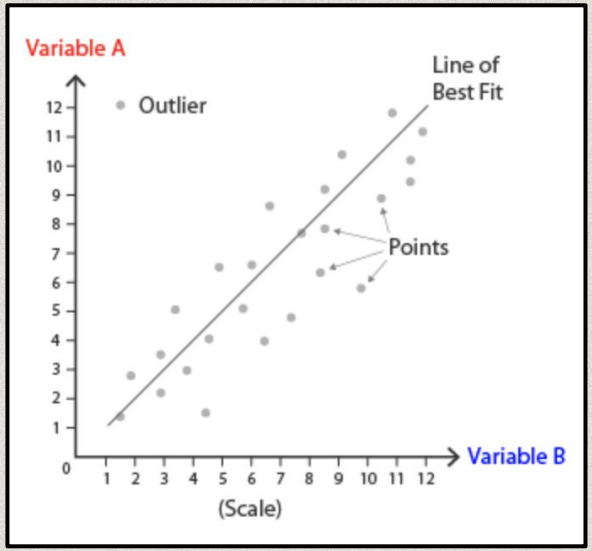
Charts review

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.



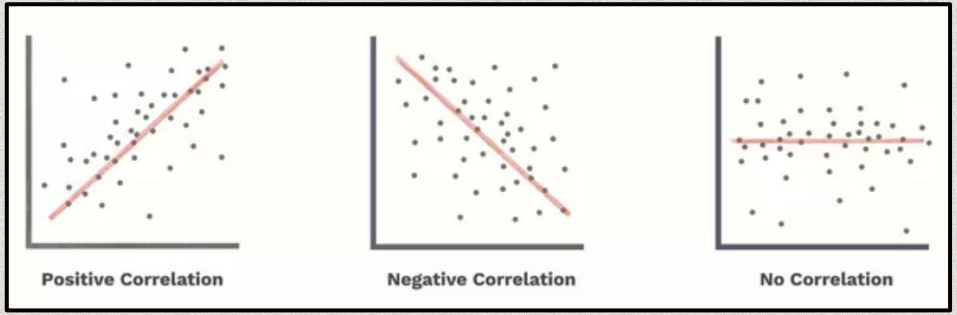
Graphs

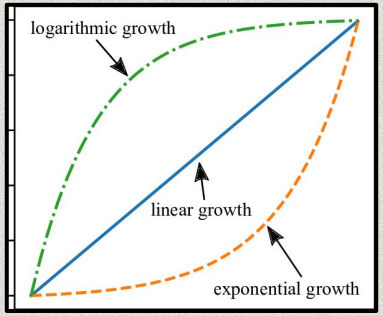
02

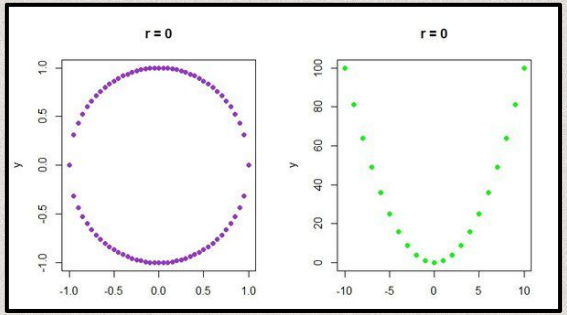
Charts review

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.





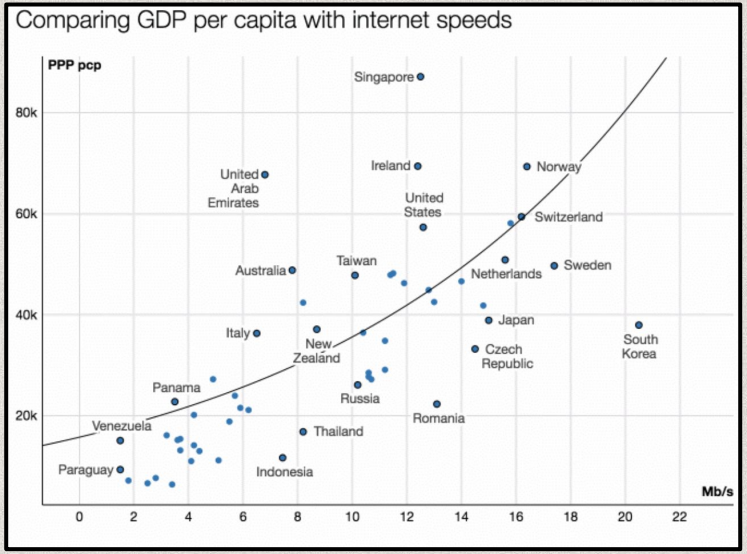


Example

Charts review

Positive correlation

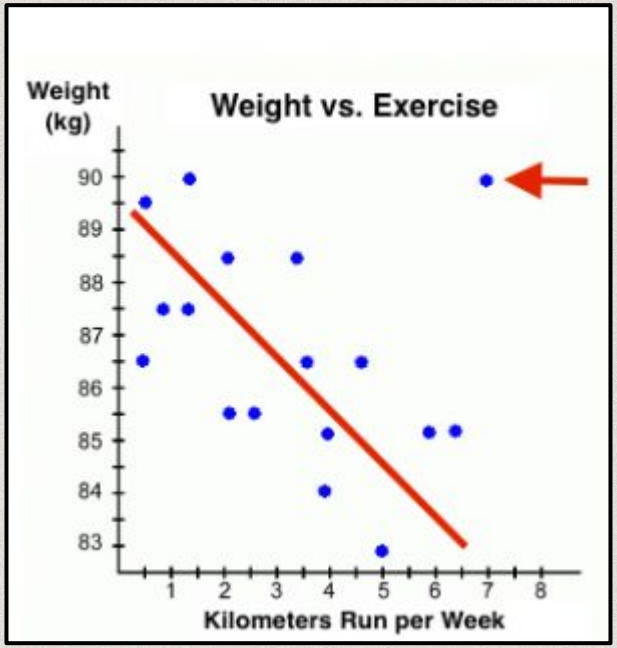
The more the income of a country grows (y axis), the more the speed of wifi grows (x axis) - and vice versa.



Example

Negative correlation

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



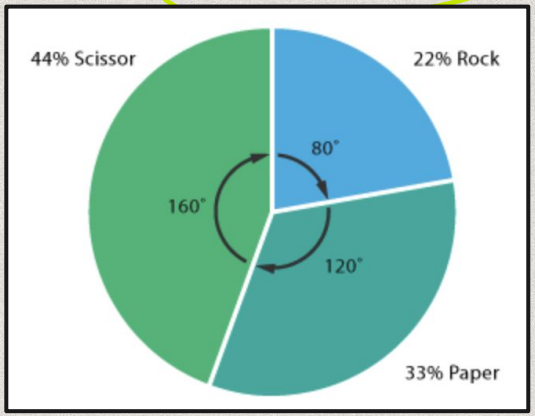
Charts

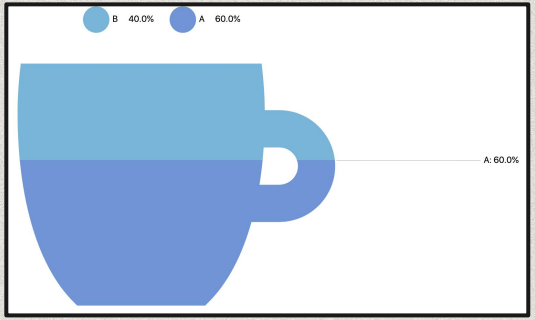
Charts review

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.





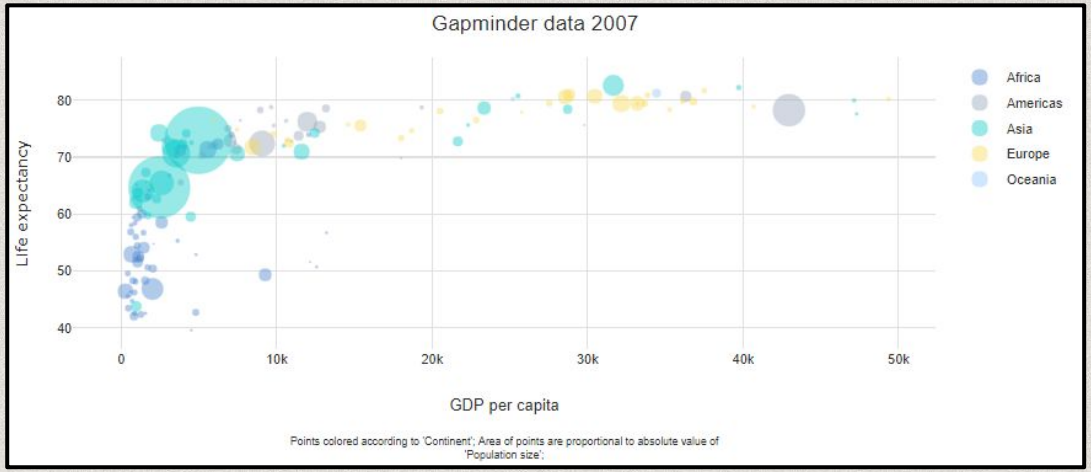
Charts

02

Charts review

Bubble chart

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

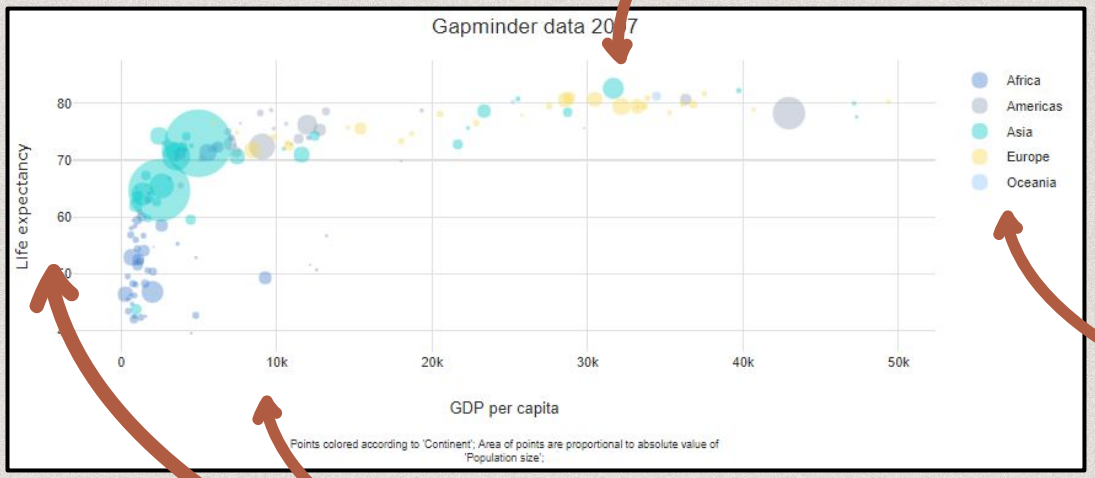


Charts

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

02

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.

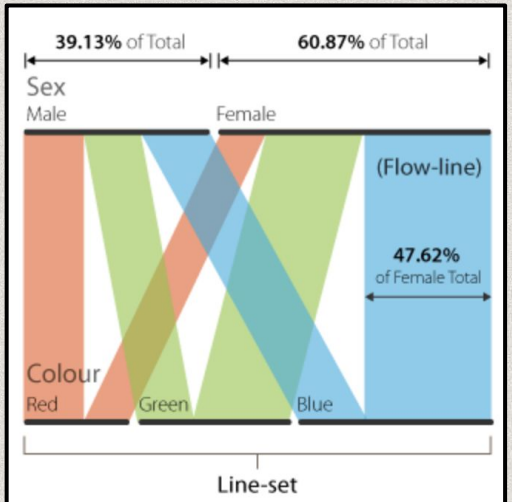
Tag clouds can be **misleading** (long words, letters with ascending lines seem bigger) and do not allow precise comparison.

Charts

Parallel sets

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

Relations are shown as flows between pairs of values.



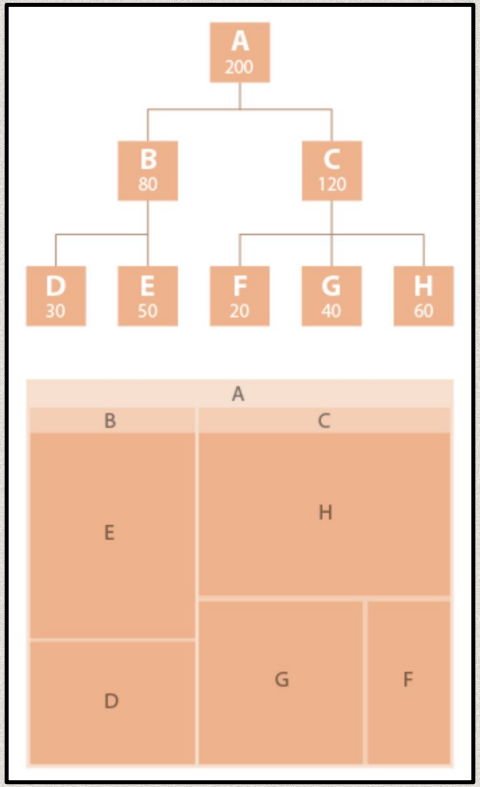
Sex	Colour	Count	% of Sex TOTAL
Male	Red	35	32.41
	Green	33	30.56
	Blue	40	37.04
Female	Red	28	16.67
	Green	60	35.71
	Blue	80	47.62

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**.

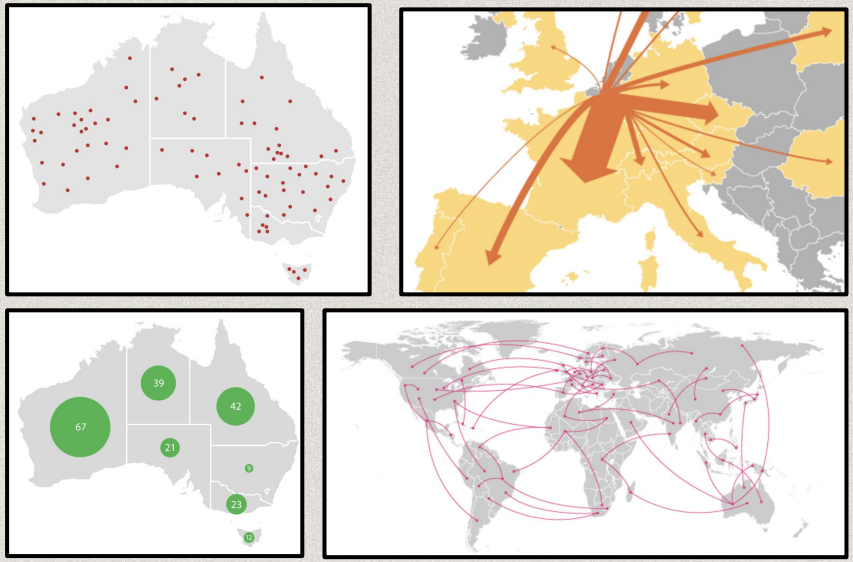


Maps

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

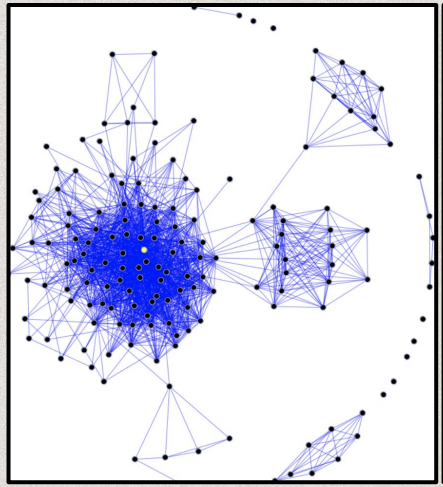
Charts review

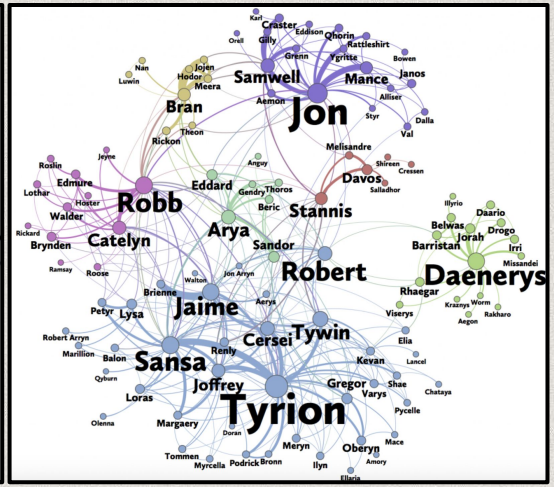
Graph networks

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.

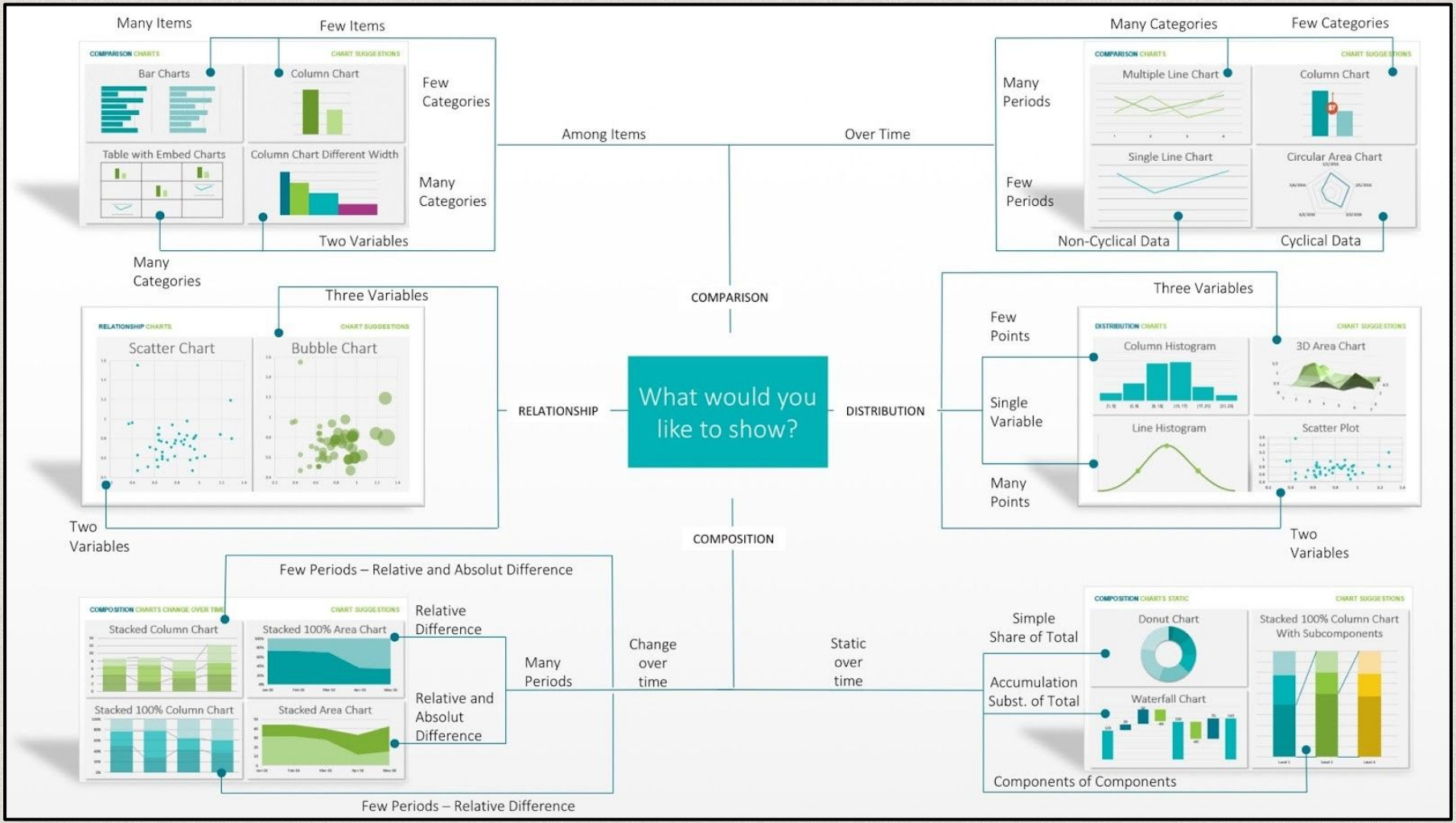
They are unreadable :(





02

How to decide?



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

Discuss

02

How to decide?

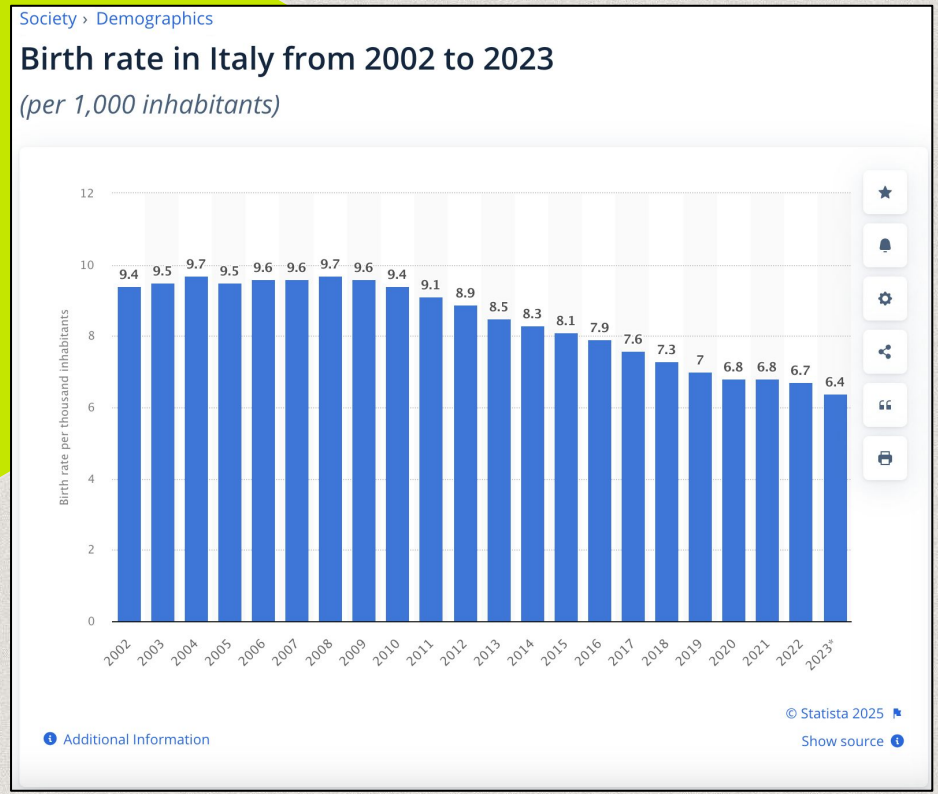
Some scenarios

1. The more classes a student misses, the more likely their grades are to decrease.
2. The number of times you look at your phone over the day.
3. The proportion of fantasy books on your shelf.

Which pattern would you highlight?
Which chart would you choose?

Discuss

How to decide?



Why do you think it happens? What other features would you explore to understand this pattern?

Discuss

02

How to decide?

Italian mothers older and older

Similar to citizens of other European countries, Italians also postpone parenthood to a later age. While the average age of an Italian mother at childbirth in the 1990s was 29.9 years, in 2022 females giving birth were roughly 32.4 years.

Italy, a country with one of the lowest fertility rates in the world

If compared with the fertility rates around the world, Italy was one of the 20 countries which registered the lowest fertility rate in 2023. The leader of the global ranking was Taiwan, where only 1.09 babies were born per woman.

Do two hints make an evidence?

TO DO

Next week

Come prepared

- Bring your laptop if possible
- Install Python3 and a IDE (e.g. PyCharm)
- Know how to run a py script via shell (read here)
- Know how to install a Python library with pip (install RDFlib)

Alternatively:

- Learn how to access and create a notebook on Colab, and how to install/import libraries there.

Thanks!

Do you have any questions?

marilena.daquino2@unibo.it

https://github.com/marilenadaquino/information_visualization

CREDITS: This presentation template was created by **Slidesgo,** and includes icons by **Flaticon,** and infographics & images by **Freepik**