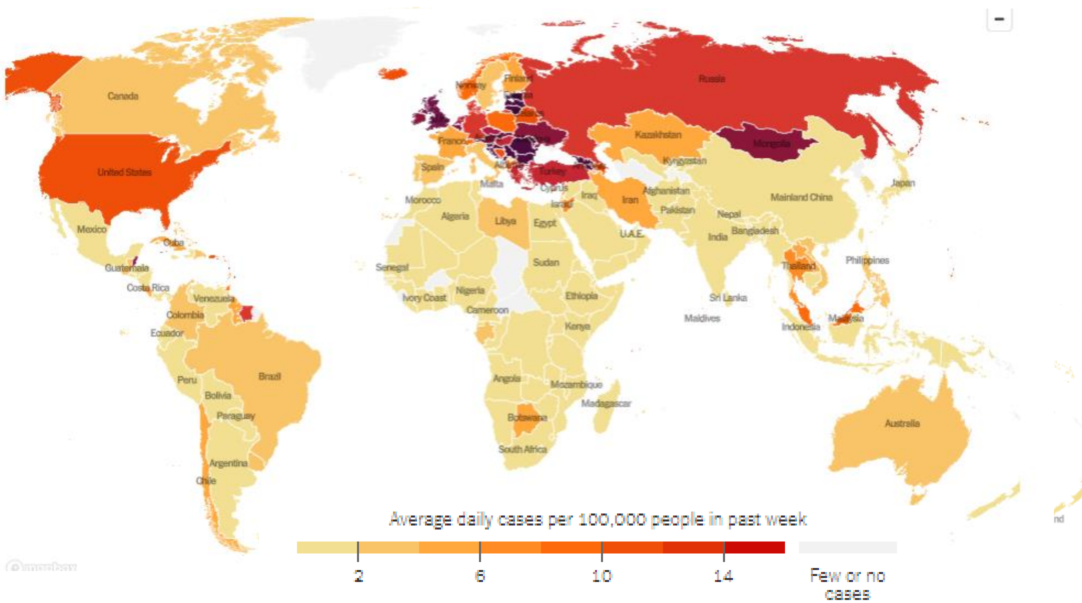
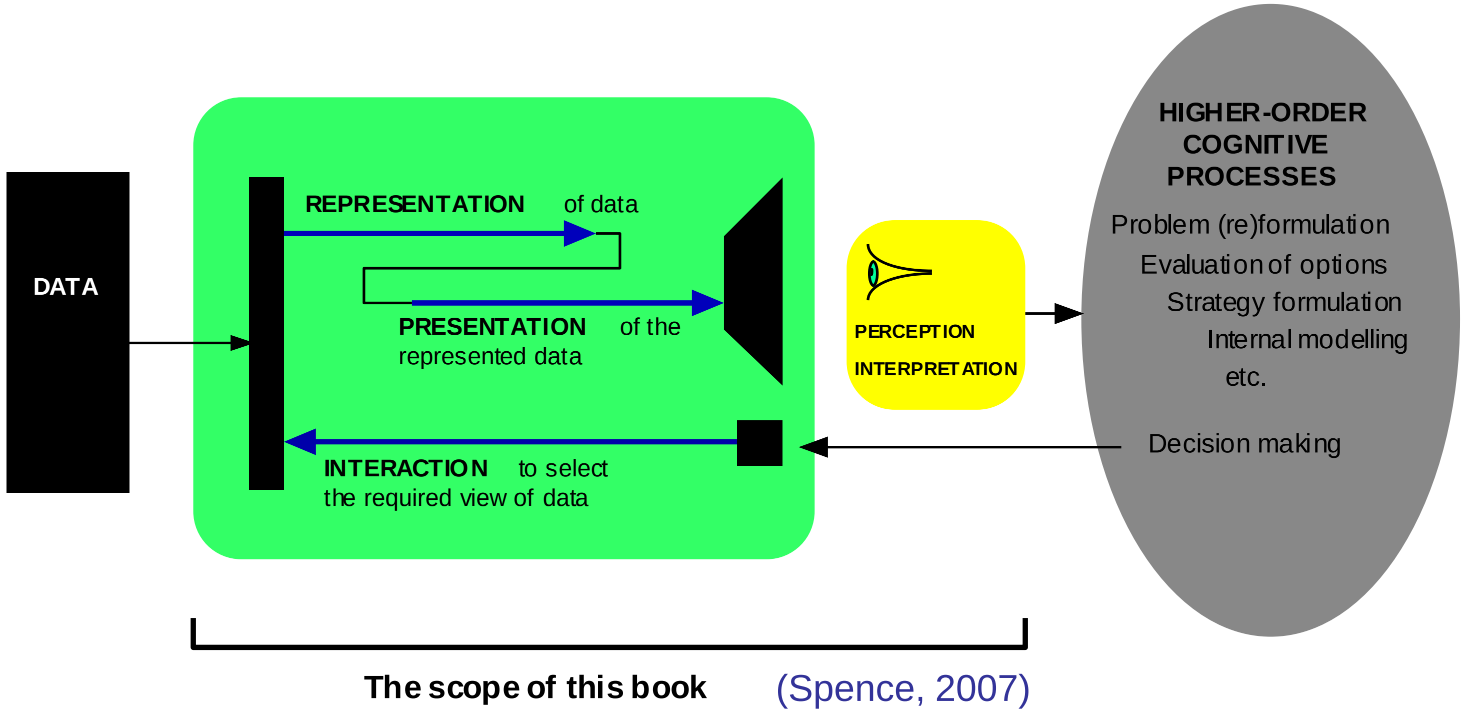
Universidade de Aveiro
Departamento de Electrónica,
Telecomunicações e Informática

Representation



https://www.nytimes.com/interactive/2021/world/covid-cases.html

Beatriz Sousa Santos, University of Aveiro, 2023

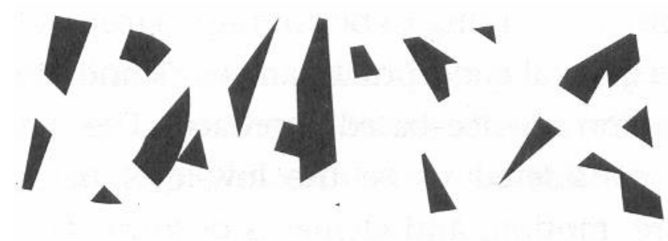


Interaction with data governed by high-order cognitive processes:

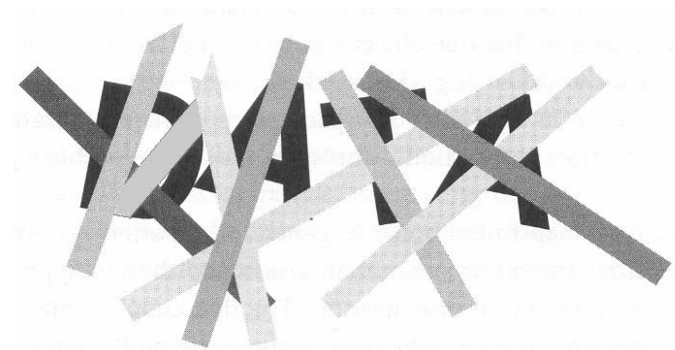
- **Representation** (how to code visually the data)
- Presentation (what/when/where to show on the screen)
- Interaction (how to let users explore the data)

Remember:

- The Human Visual system is the product of millions of years of evolution



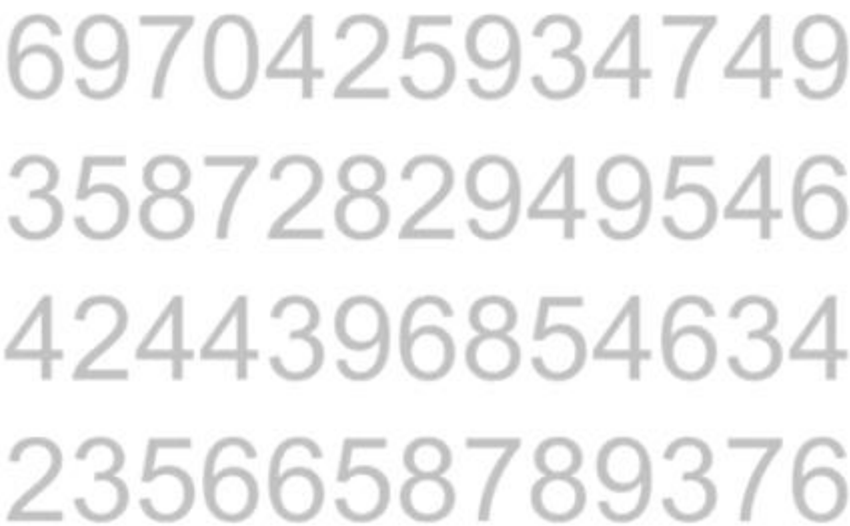
- Although very flexible, it is tuned to data represented in specific ways



- If we understand how its mechanisms work we will be able to produce better results

Pre-attentive attributes can help observers to see before though

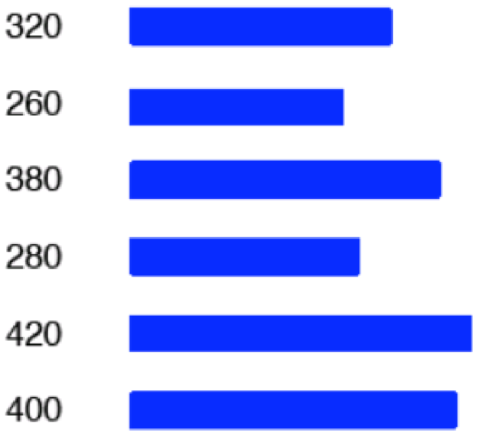
Example: Count the number of 7s



https://www.youtube.com/watch?time_continue=121&v=AiD6etOB6qI

- Visual attributes as **size**, **proximity** are quickly processed by visual perception, **before the cognitive processes** come into play

Example: mapping numerical values to the length of bars:



(Mazza, 2009)

Designing a Visualization

- The process must be preceded by good design

- The main problem in designing a visual representation is the choice of mapping, as to:

  - help the user to attain their goals

  - faithfully reproducing the information codified in the data

- The visual representation suitable depends on:

  - the nature of the data and phenomenon
  - the users' tasks and needs (the questions)
  - the user profile and context of use
  - …

Procedure to follow to create visual representations of abstract data

Taking into consideration **the users' tasks, profile and context of use**:

1. Define the problem and the **users' questions, profile and context of use**
2. Examine the **nature of the phenomenon and data**
3. **Pre-process** the data
4. Determine the **number of attributes**
5. Choose the **visual structures** (how to represent the data)
6. Establish **how and when** to present to the user

   and the type of i**nteraction**

test several ideas and redesign when needed…

nature of the problem
- communicate
- explore
- confirm

nature of the data to represent
- quantitative
- ordinal
- categorical

number of attributes
- univariate
- bivariate
- trivariate
- multivariate

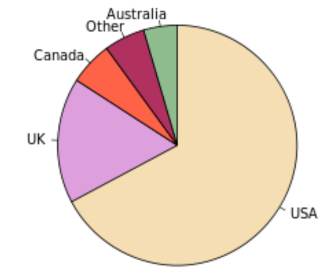
Next: representation methods organized according the n. of attributes

data structures
- linear
- temporal
- spatial or geographical
- hierarchical
- network

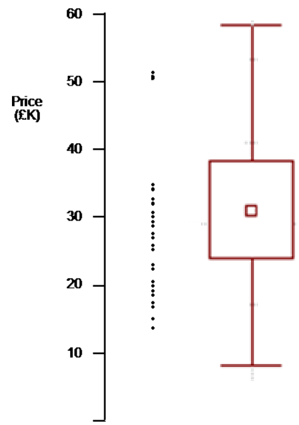
,

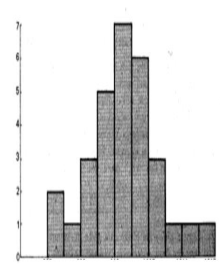
type of interaction
- static
- transformable
- manipulable

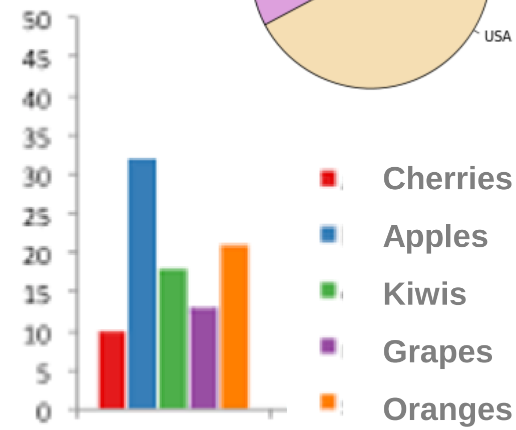
Common Visualization Techniques
for **univariate**, **bivariate and trivariate** data



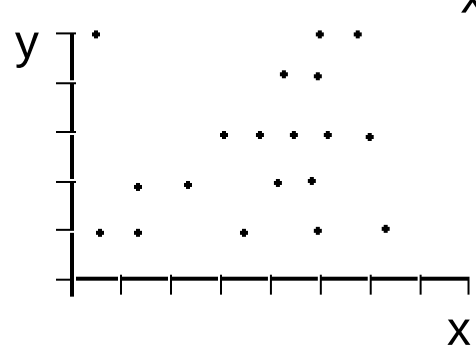
Univariate data
- dot plot
- box plot
- bar chart
- histogram
- pie chart
- …

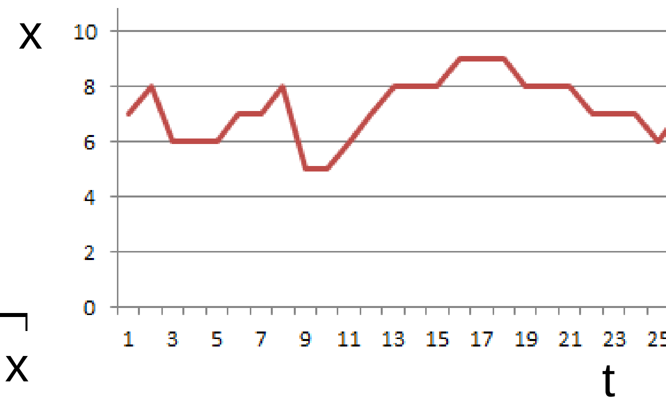




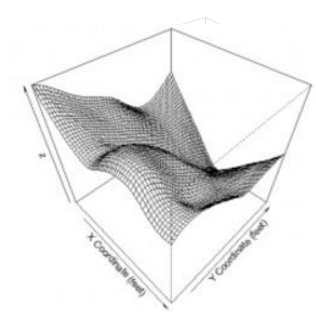


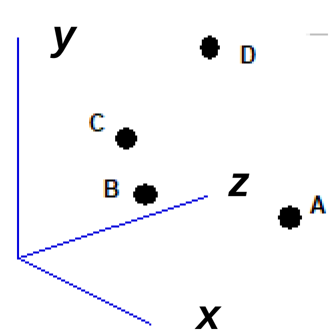
Bivariate data
- scatter plot
- line plot
- time series
- …

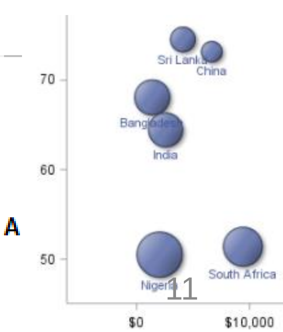




Trivariate data
- surface plot
- contour plot
- 3D representation
- bubble plot
- …





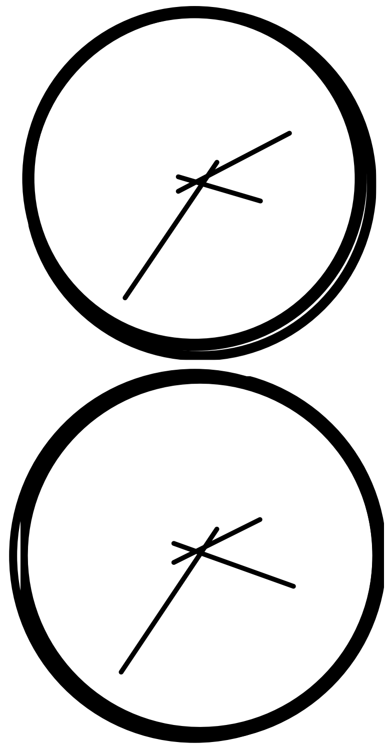


Representing univariate data

- A **single number** can be difficult to represent ensuring a user is made aware of it

Example: the altimeter
(Spence, 2007)

The original type of aircraft altimeter, with usability issues



Two altimeter representations easily assumed to be the same due to change blindness

A more usable solution for altimeter display

Some more examples on how humans see…

And their limitations…

Example of change blindness
(Spence, 2007)

Example of change blindness
(Spence, 2007)

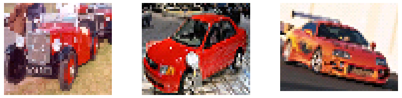
What is missing now?

Inattentional blindness
https://www.youtube.com/watch?v=IGQmdoK_ZfY

Change blindness
http://www.youtube.com/watch?v=vBPG_OBgTWg&feature=related

Representing univariate data (cont.)

- A more common situation consists in representing a set of values
- Well established techniques exist
- But new ones can be invented!

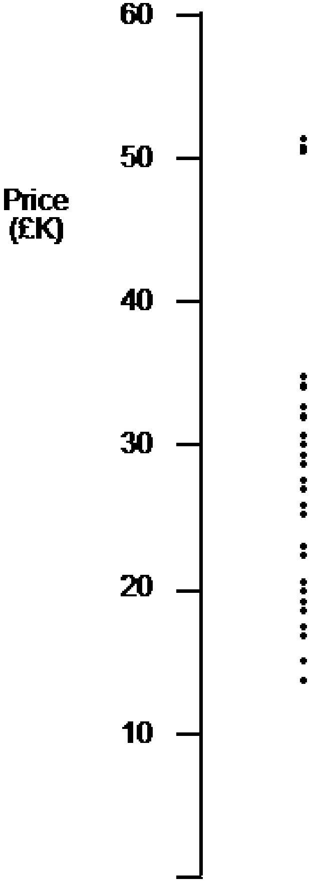
Example: 

Price for a number of cars:
- dots on a linear scale
- box plot

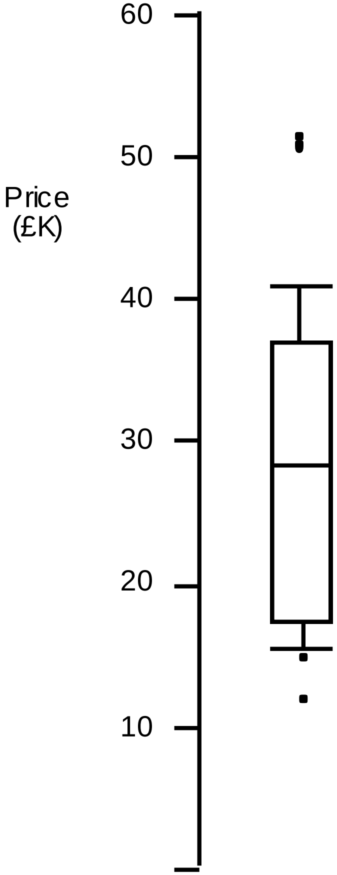
(that will answer many questions: median value, outliers,...)

(Spence, 2007)

Dot plot

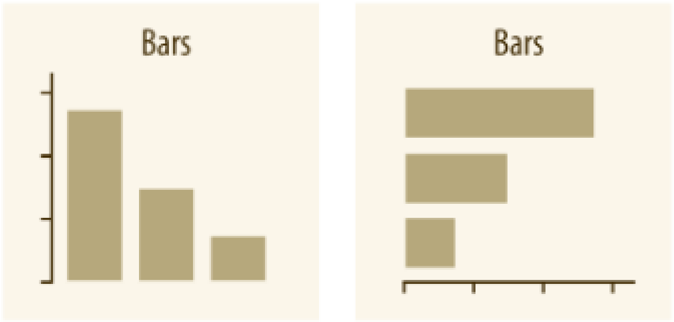


Tukey boxplot

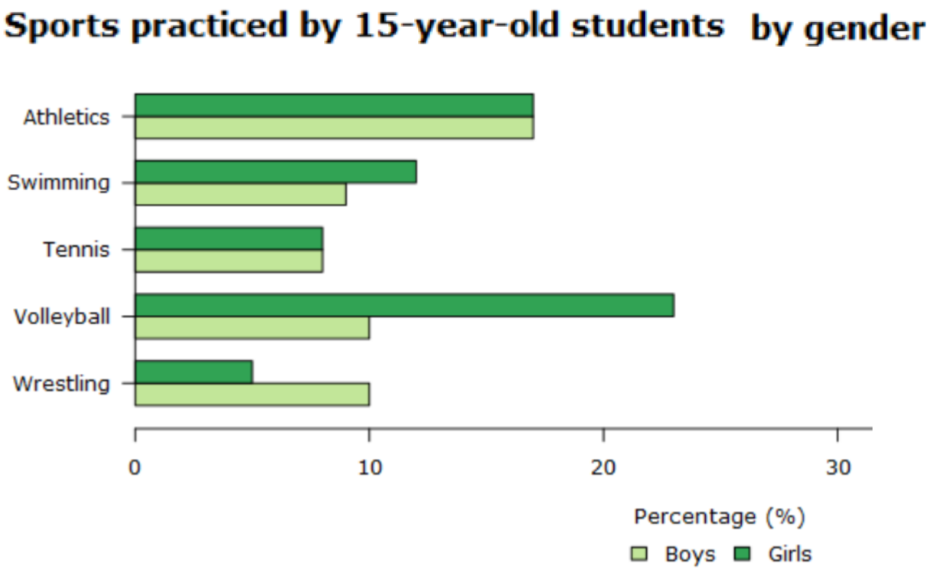


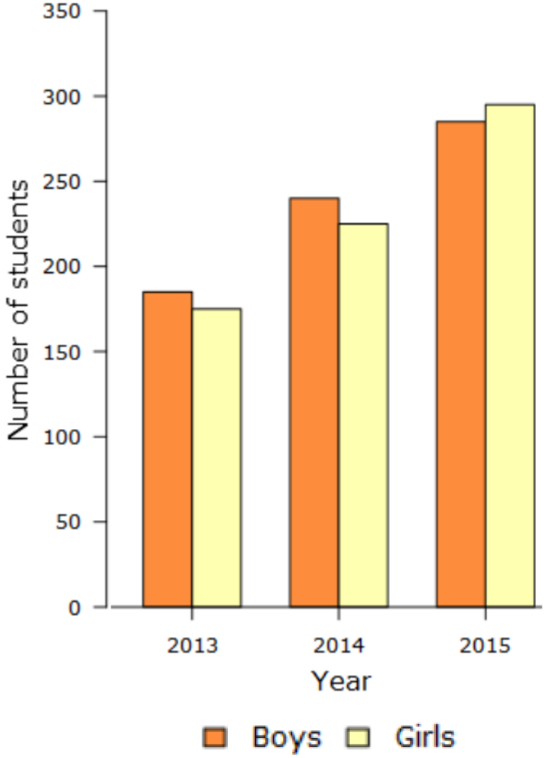
https://www.data-to-viz.com/caveat/boxplot.html

- much of the data is aggregated
- precise detail is often not needed
- A bar chart is a common way to represent one attribute,
- but we may combine to represent more attributes



(Wilke, 2019)





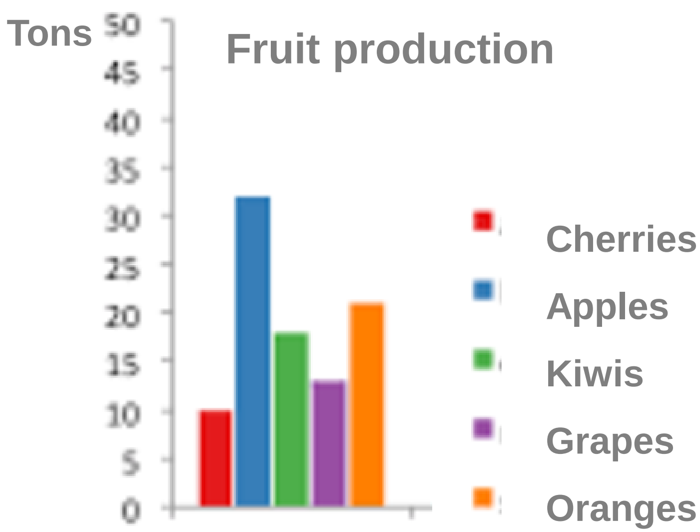
Simple (and common) representations of one attribute data

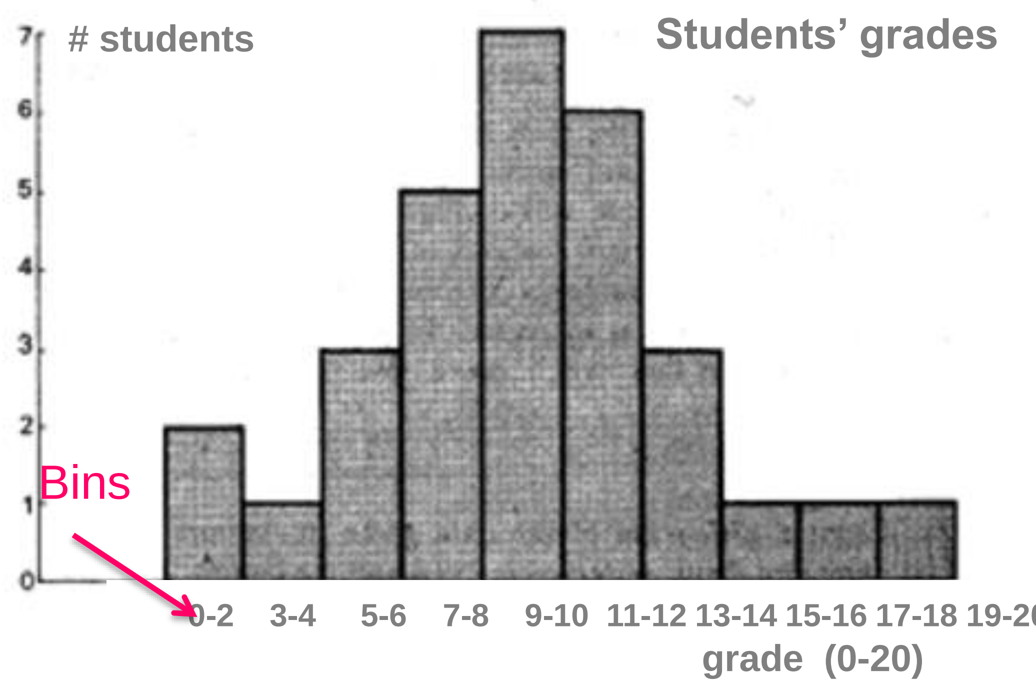
- Two common techniques not to be confused !

Histogram → represents a distribution of numerical data

Bar chart → represents number of occurrences of categorical/ordinal data

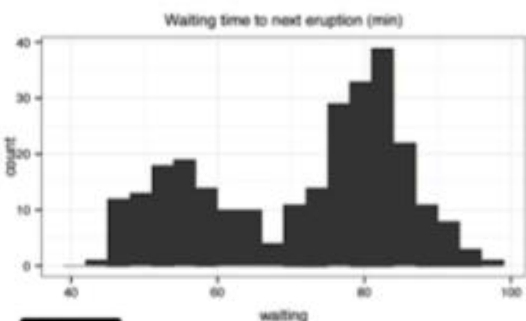
Both represent data by rectangular bars(vertical or horizontal) with length proportional to the values they represent

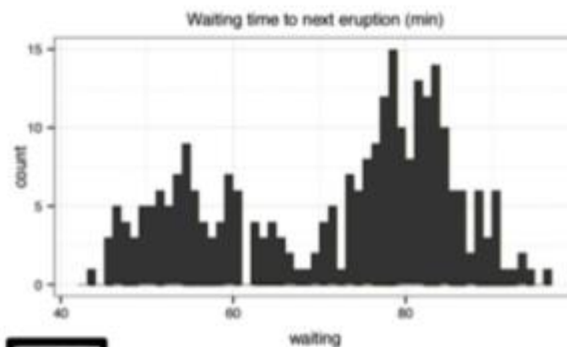


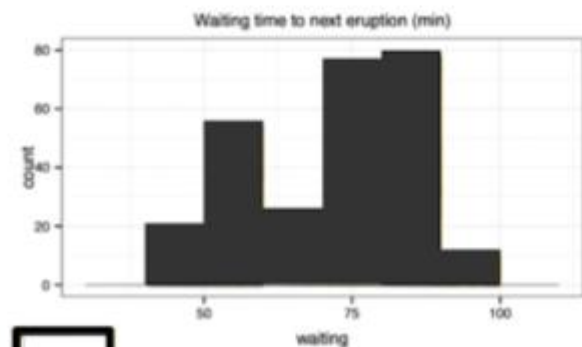


Histogram Quiz

Given the following plots with different bin widths, **Match the description to the plot.**





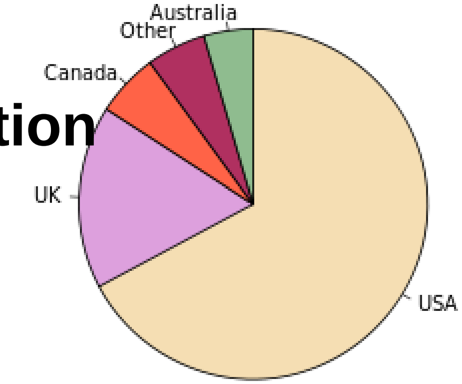


A: good bin width - shows important signal in data (two modes) but not too much noise.

B: bin width is too small

C: bin width is too big

Another simple (and too common) representation



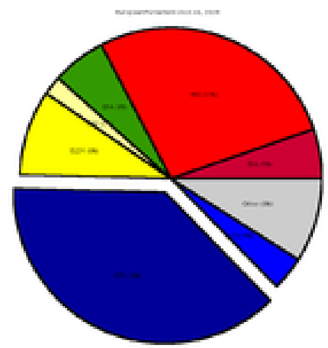
Native English speaking population

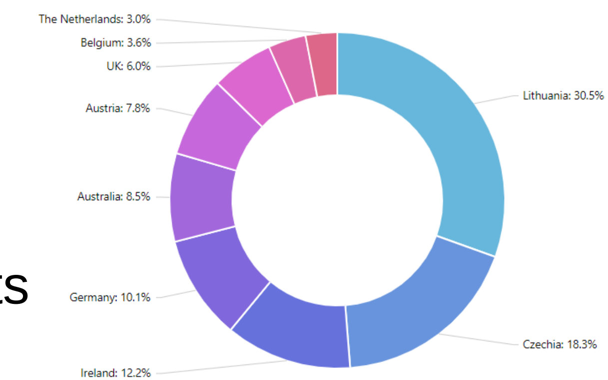
- Pie Chart

Represents numerical proportion, **parts of an whole**

The arc length of each slice (its central angle and area), is proportional to the quantity it represents

Variations of pie charts:





Are much controversial:

many experts recommend avoiding them

http://www.perceptualedge.com/articles/08-21-07.pdf

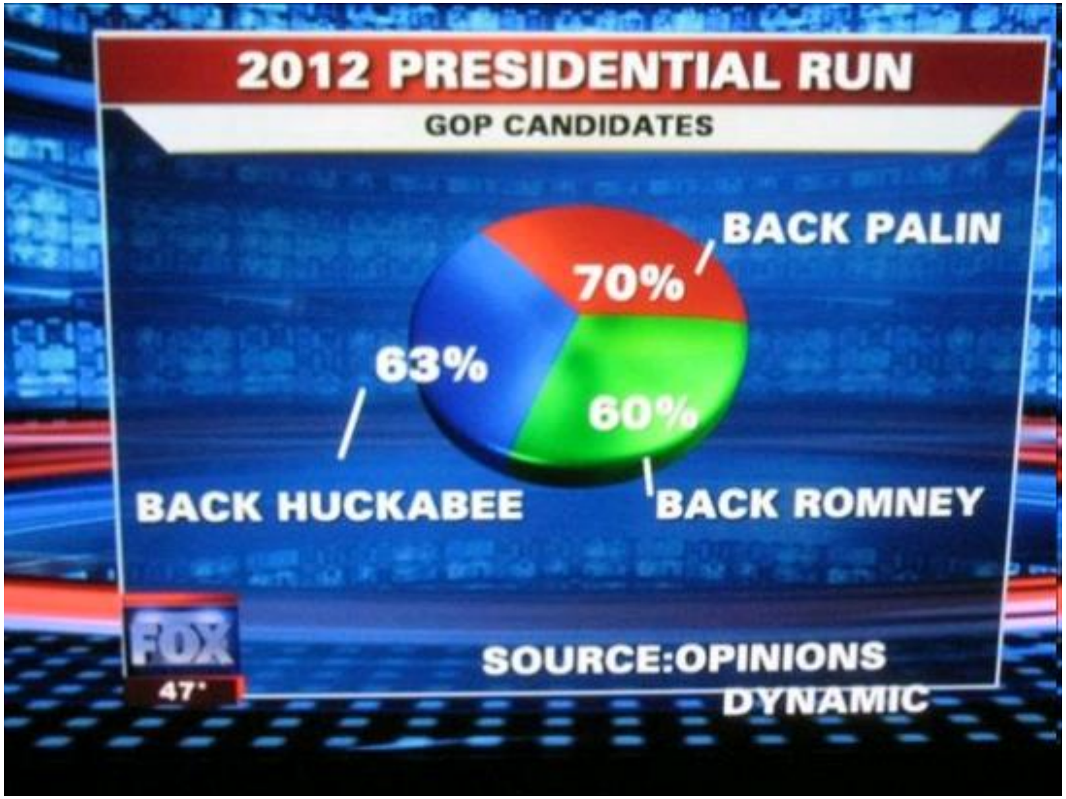
It is difficult to compare different sections of a pie chart, or to compare data across different pie charts

- Simple criteria to determine whether a pie chart is acceptable
- Consider it **only if**:
- **The parts make up a meaningful whole**
- **The parts are mutually exclusive**
- **There are <6 parts and slices have not very different sizes**

If the main purpose is to compare between the parts,
use a different chart!

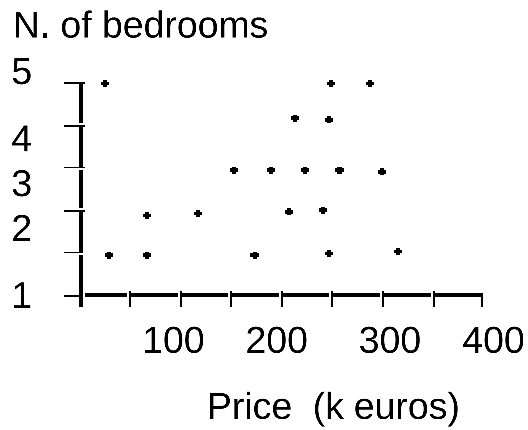
https://eagereyes.org/techniques/pie-charts

Prize winning Pie chart!



Representing bivariate data

- The **scatterplot** is a conventional representation



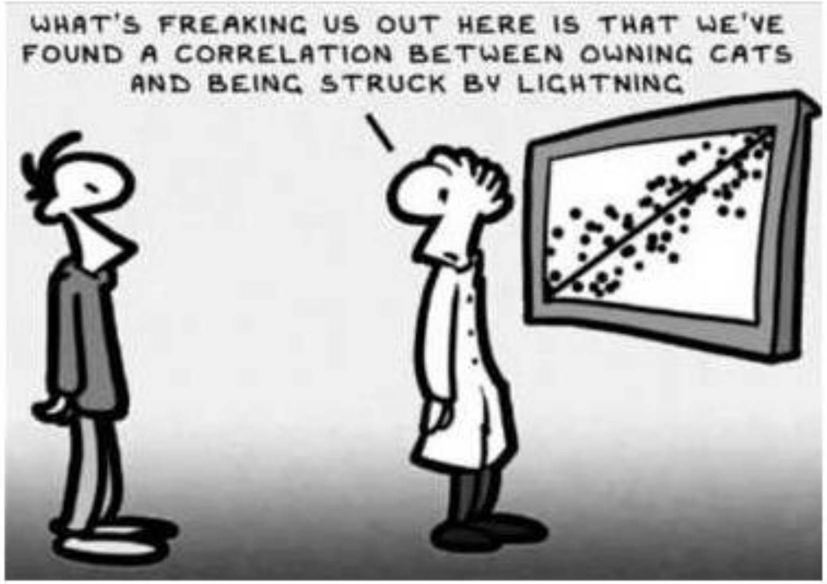
Each observation is represented by a point on a two dimensional space

The axes are associated with these two attributes

This representation affords awareness of:

- general trends
- local trade-offs
- outliers

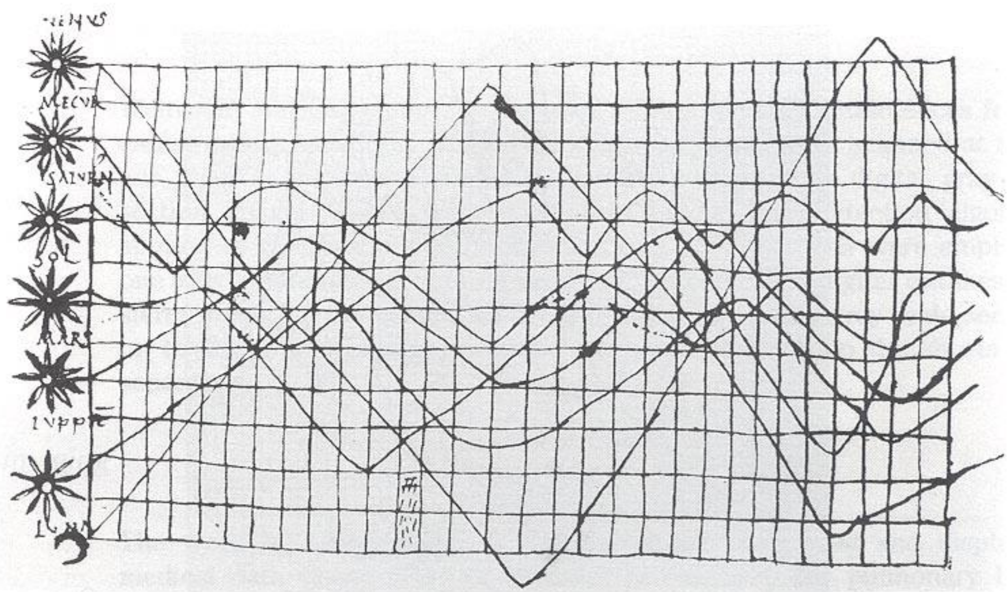
Correlation is not causation



Representing bivariate data

The **line chart**

One of the oldest known and ubiquitous Visualizations

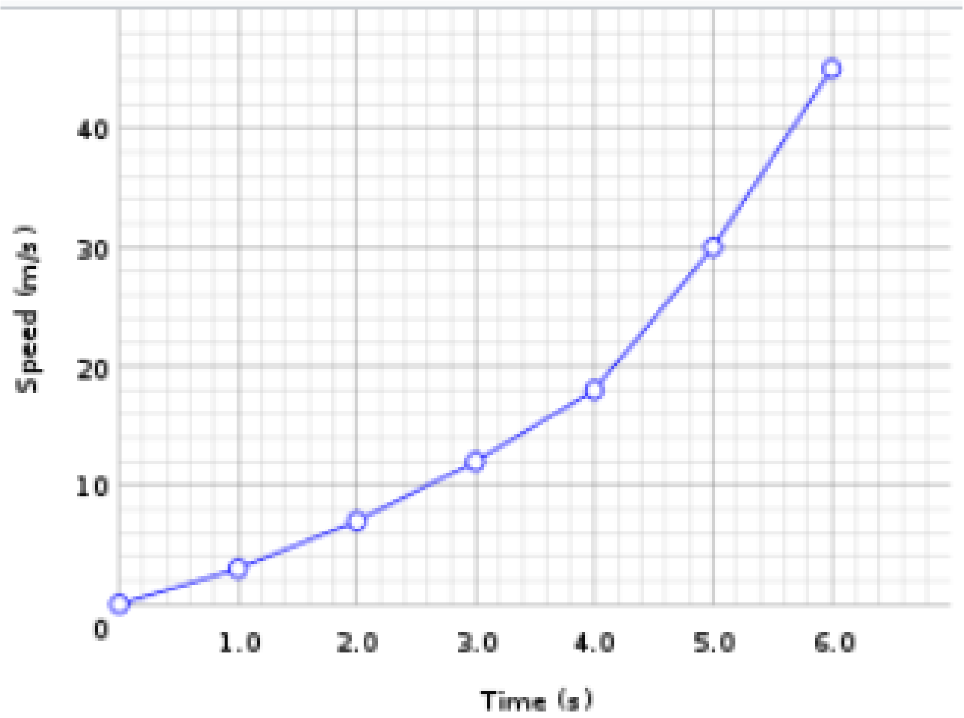


Inclination of orbits along the time - Xth century (Tufte,1983)

- A **line chart** or **line plot** or **line graph** or **curve chart** displays information as a series of data points called 'markers' connected by straight line segments

- Basic type of chart common in many fields

- Often used to visualize a trend in data over intervals of time



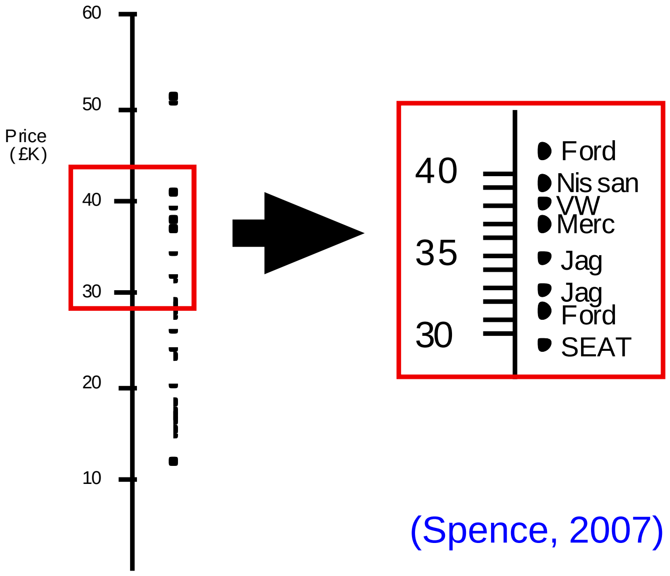
- If one attribute is more important than the other or must be examined first,
- it may be appropriate to employ logical or semantic zoom

Example:

Analyzing a list of cars:

- price is the first attribute to examine

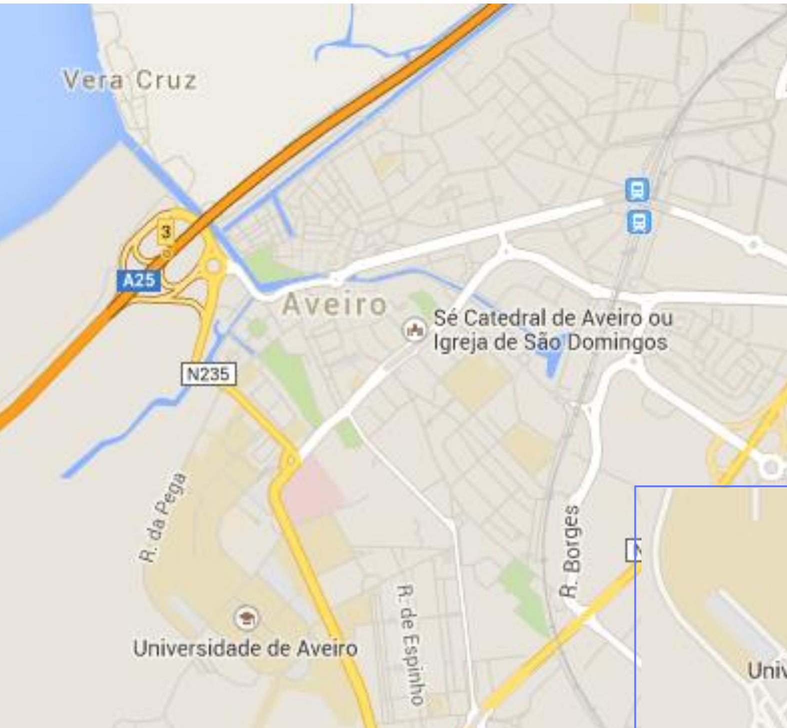
- semantic zoom reveals data about
  a second attribute



(Spence, 2007)

- This technique is quite general: it can encompass many attributes and many levels of progressive zoom

Example: Zoom in Google Maps

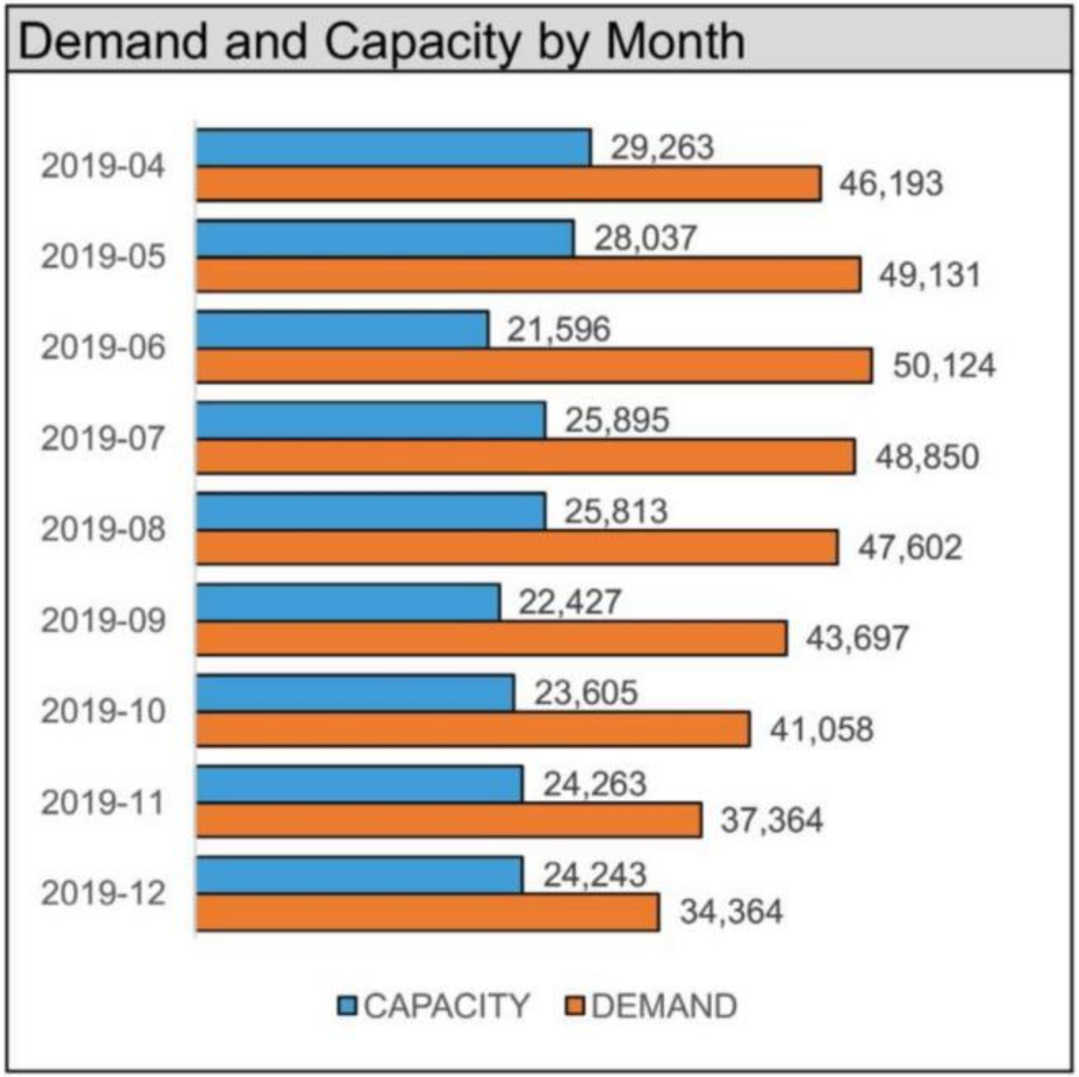


Analyze this visualization

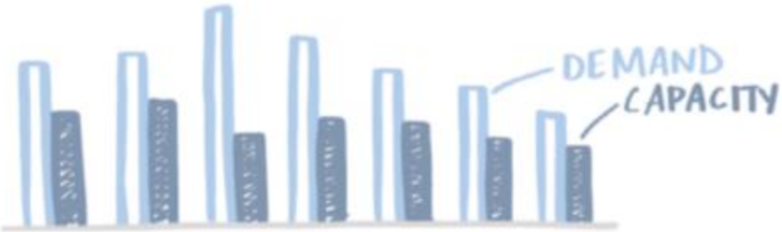
How many attributes?

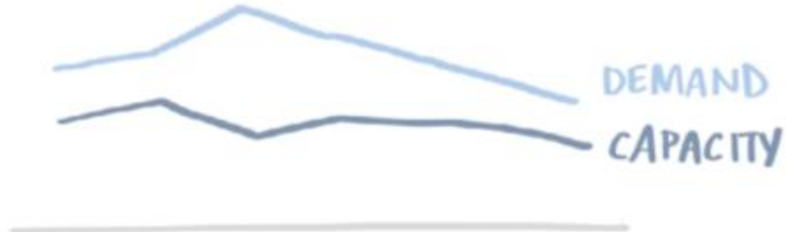
What type of attributes?

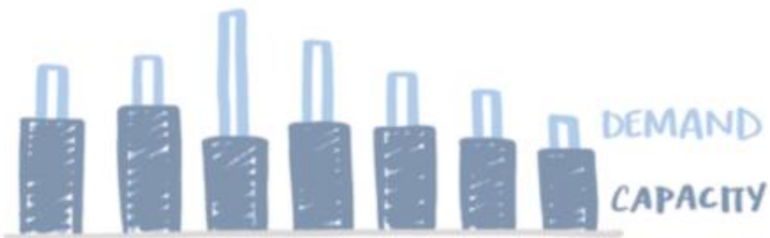
Can you think of other ways of visualizing this dataset?



Other possibilities:
what are the advantages/disadvantages of each?







Representing Trivariate data

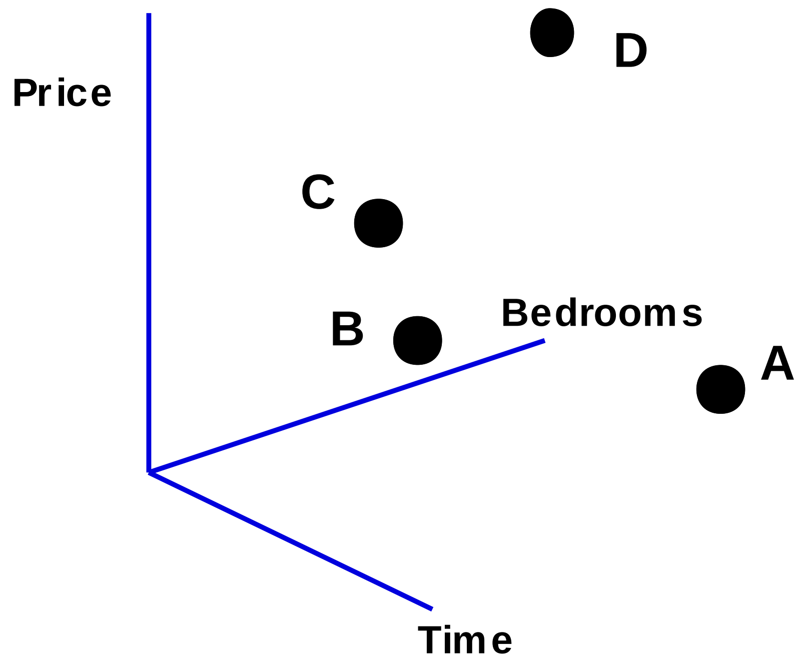
- Since we live in a 3D world, representing trivariate data as points in a 3D space and displaying a 2D view seems natural

- However, these representations can be **ambiguous**

- This can be mitigated by interaction, allowing the user to reorient the representation

Generally, avoid 3D in InfoVis!

for 3D to be useful, you' ve got to be able to move it"

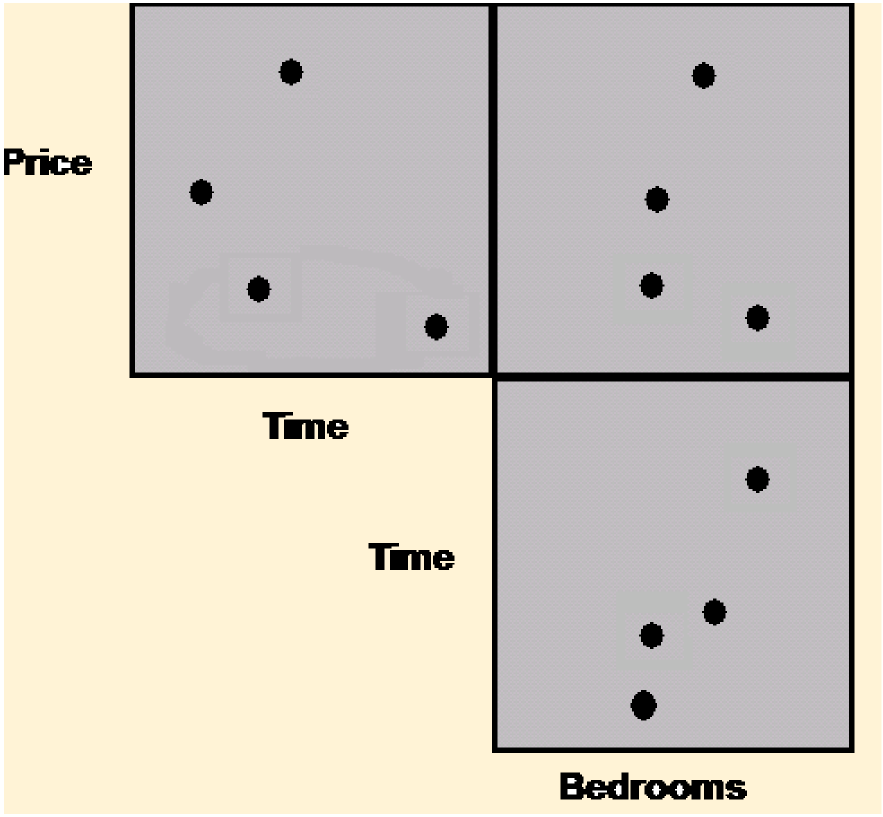


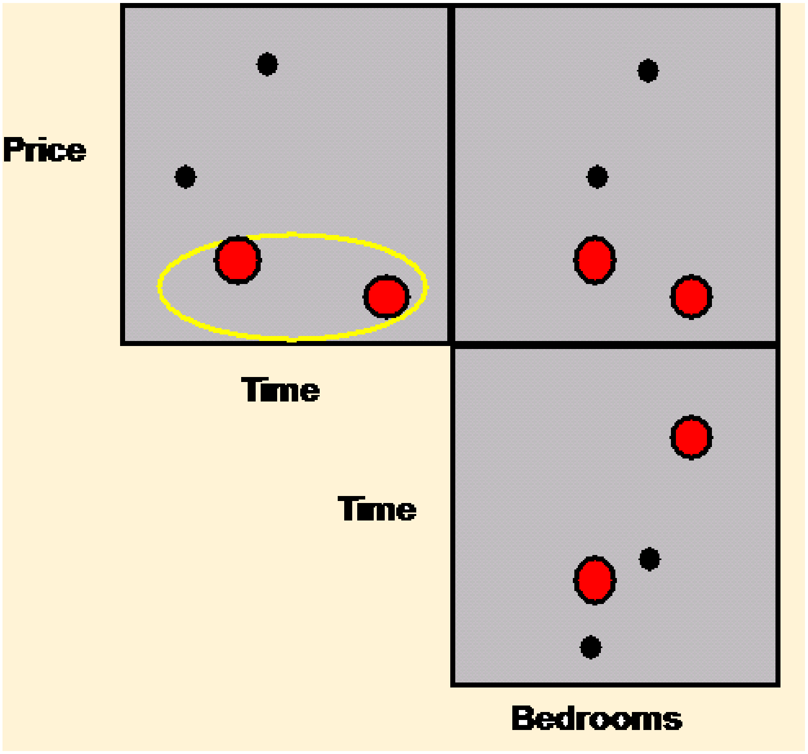
(Spence, 2007)

- Interaction (brushing) can help – objects identified in one view are highlighted in the other two planes

- change blindness must be taken into account and ensure that the user notices the highlight in the other two planes

(Spence, 2007)

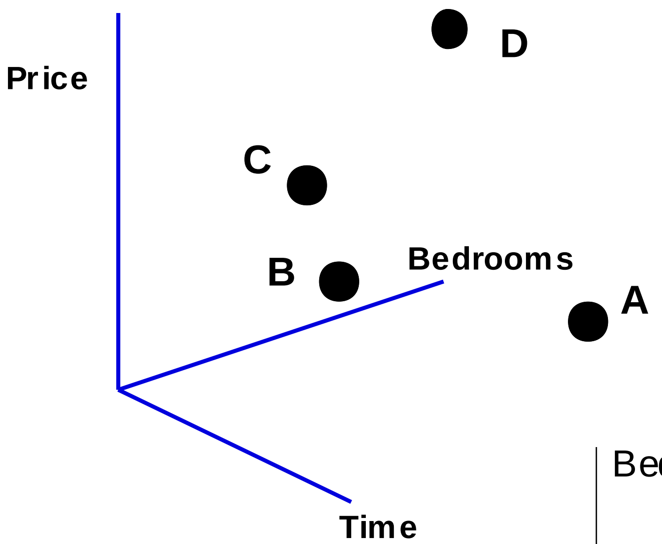




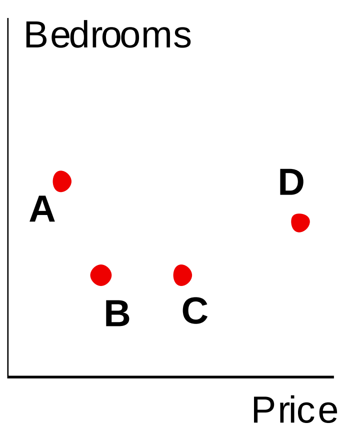
The highlighting of houses in one plane is brushed into the remaining planes.

- An alternative representation for trivariate (and hypervariate) data is a structure formed from the three possible 2D views of the data

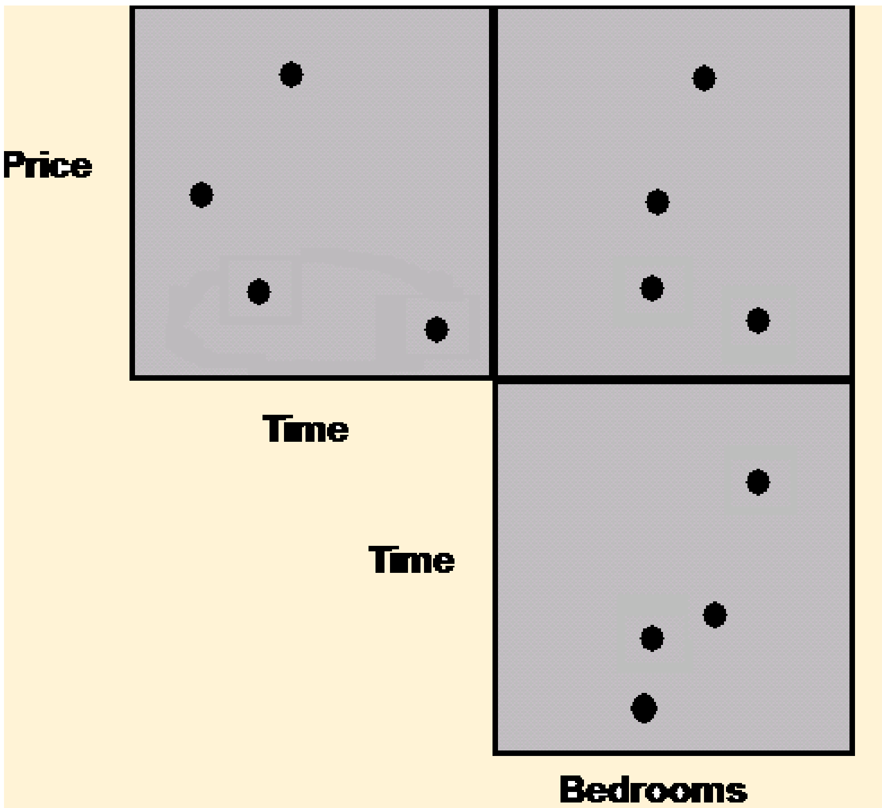
Example: houses (price, number of bedrooms, time of journey to work)



(Spence, 2007)



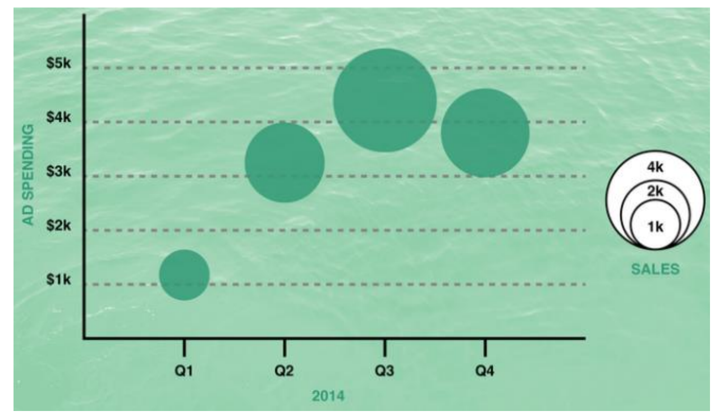
Scatterplot matrix



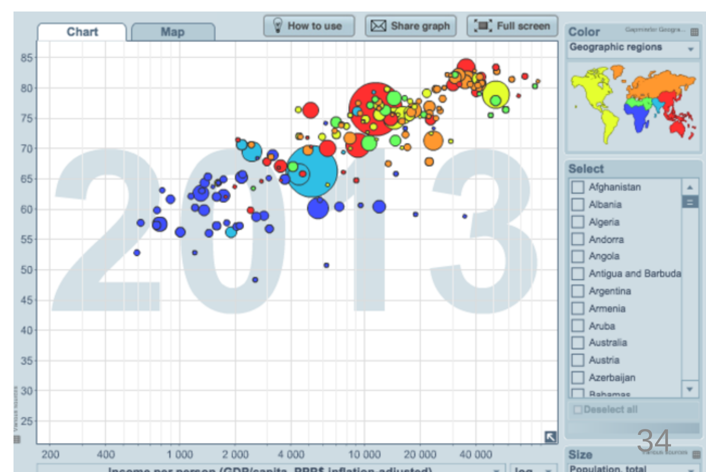
Other Simple (and common) representation of 3D data

- In a **bubble chart** data are represented as a disk that expresses two of the values through the disk's *xy* location and the third (less accurately) through the size of the disk (radius or area?)

- Mapping the atribute/variable to disk size must be done carefully. The interpretation of may be ambiguous



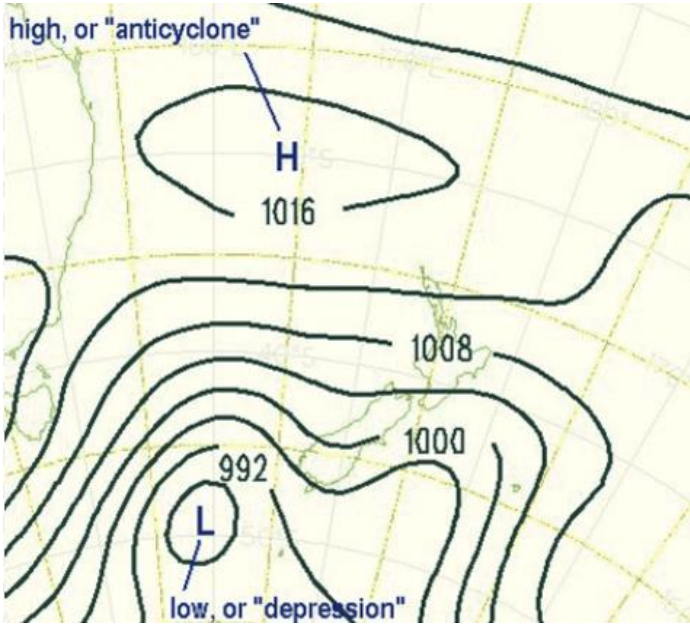
- Representing one more dimension
  through color



https://visage.co/data-visualization-101-bubble-charts/

Simple representations of a function (field) of two variables

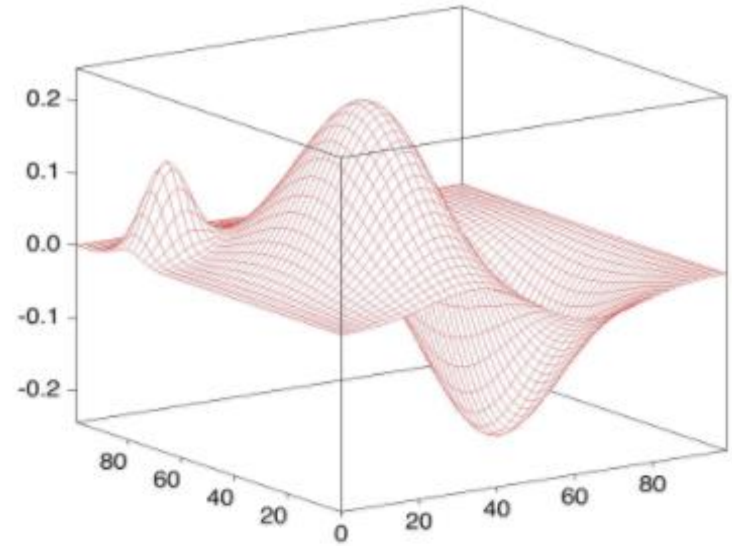
- Contour plots
- **contour line** (also **isoline**, **isopleth**, or **equipotential curve**) of a function of two variables is a curve along which the **function has a constant value**, so that the curve joins points of equal value.
- Often used in SciVis
- Typical in meteorological charts (isobars and isothermal curves)
- and maps (to represent altitude or depth)

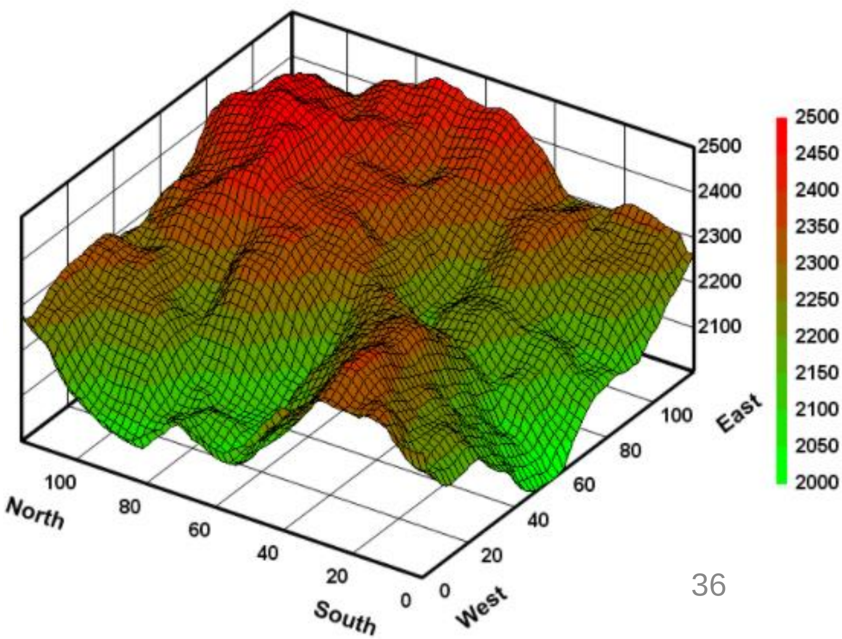


- Surface plots (also often used in SciVis)

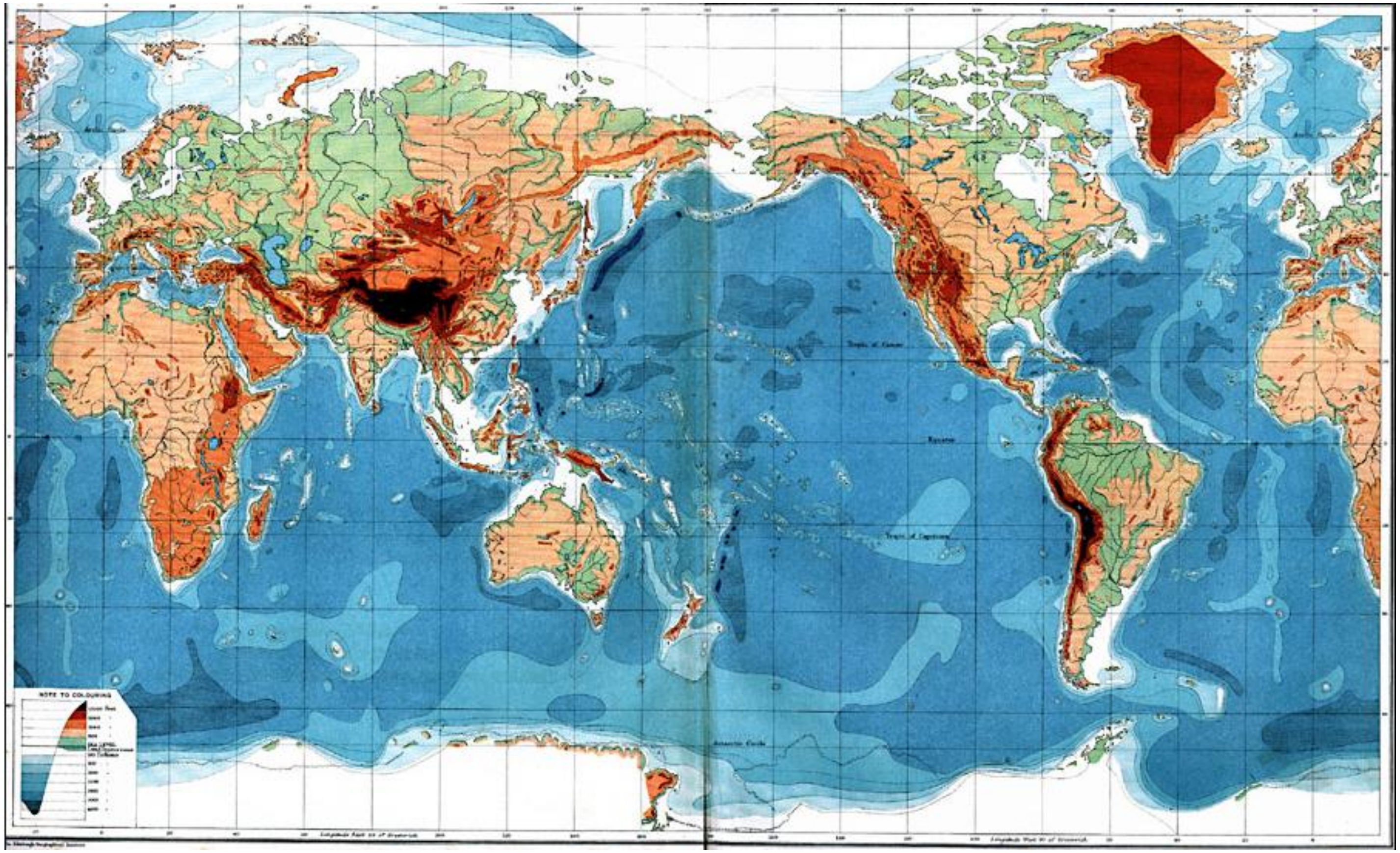
- May be combined with color

(preferably in a redundant way and carefully selecting the scale)





A special category of trivariate data:
Maps (latitude and longitude + a value)

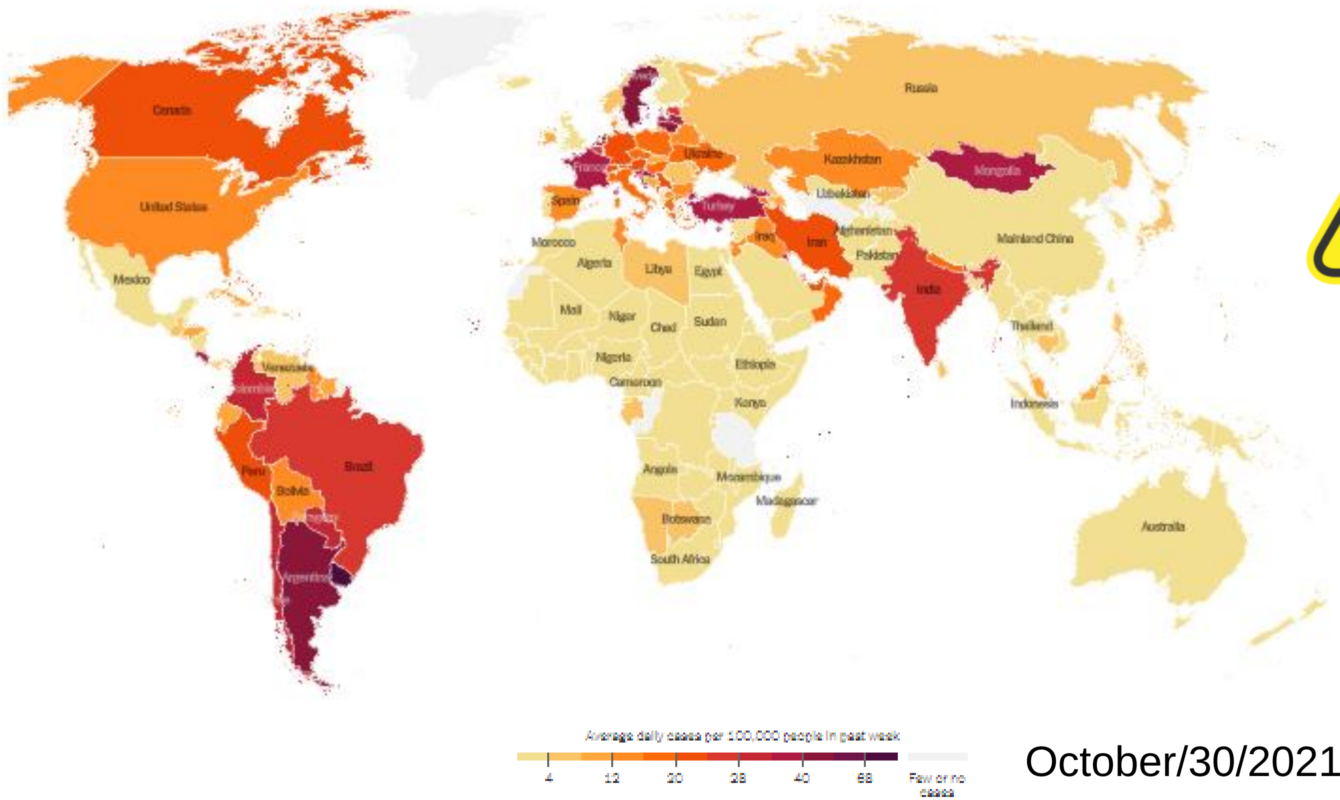


1915 – Orographic Chart of the World

https://etc.usf.edu/maps/pages/100/167/167.htm

Choropleth maps - A standard approach to communicating aggregated data by geographical areas using color encoding of the geographic area

They require some care: what are the possible issues?



October/30/2021

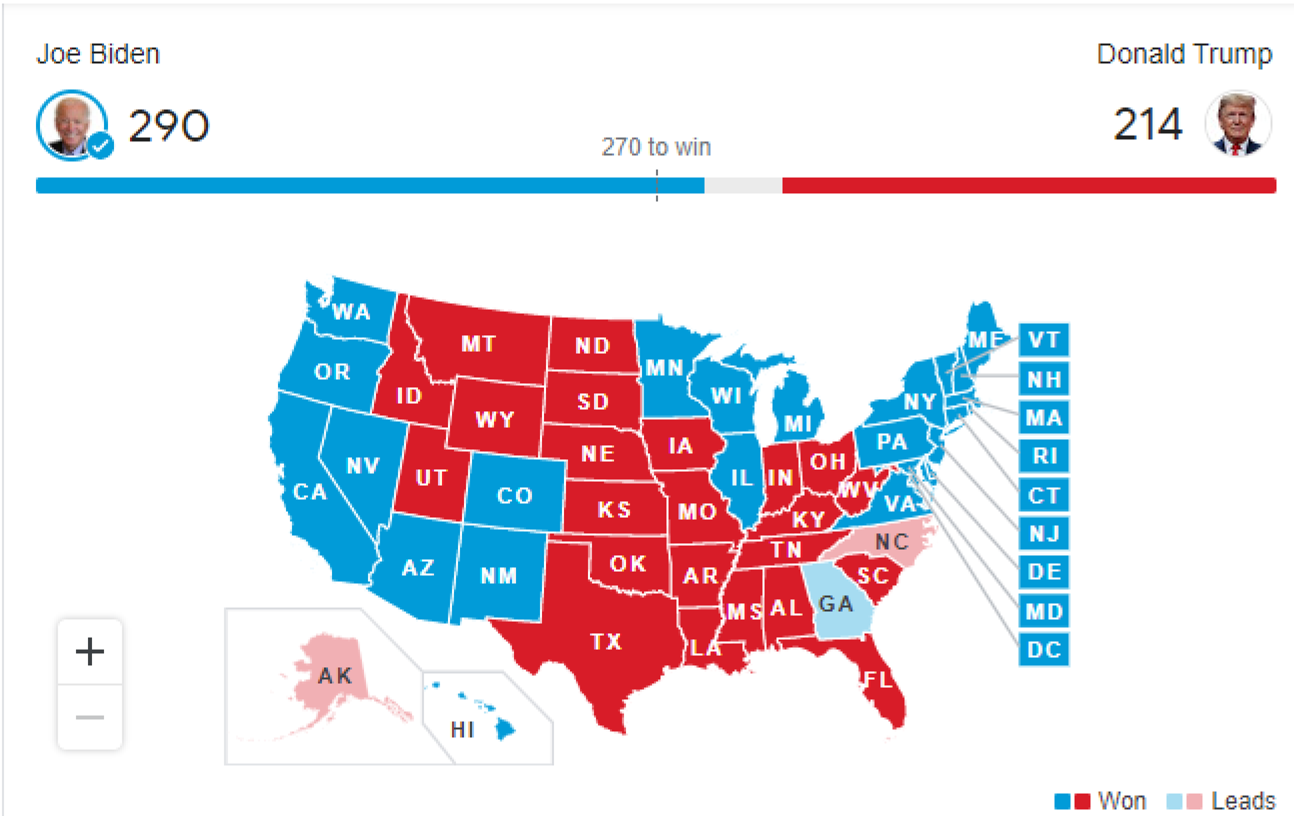
https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html

How can these issues be mitigated?

Covid vaccination worldwide
(choropleth + details on demand)

https://www.nytimes.com/interactive/2021/world/covid-vaccinations-tracker.html

Visualizations of the US 2020 Election (choropleth + bar)

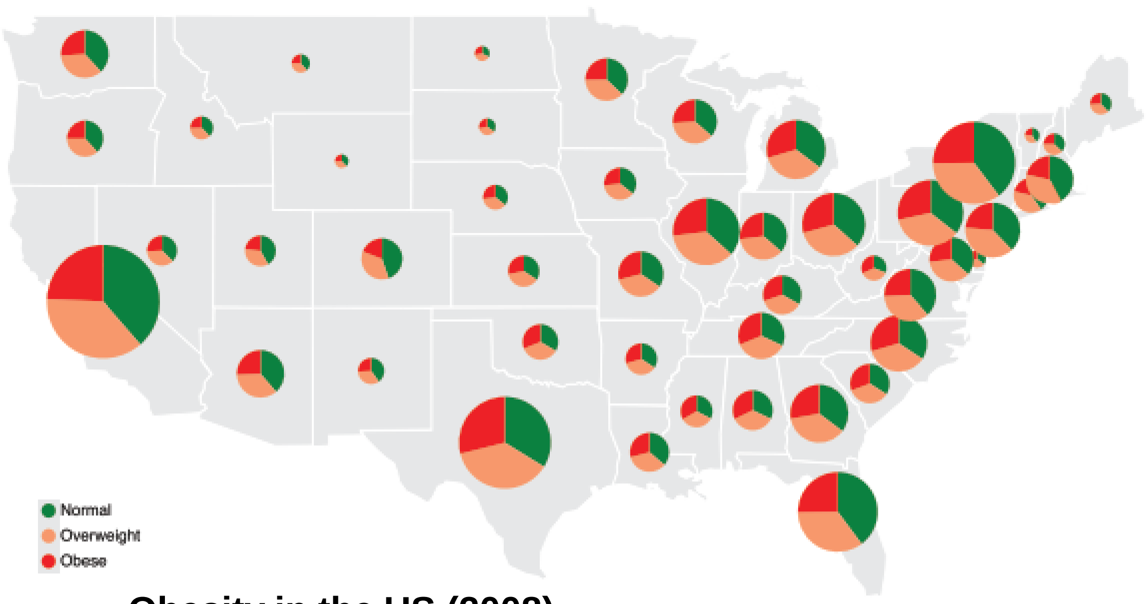


Candidates	Electoral votes	Vote %	Vote count
Joe Biden Democratic Party	290	50.6%	75,198,127
Donald Trump Republican Party	214	47.7%	70,804,457

- **Graduated Symbol Maps** are an alternative to the choropleth map;

- Symbols are placed over an underlying map; may show more dimensions

- Avoid confounding geographic area with data values



Obesity in the US (2008)
(Heer et al., 2010)

- **Cartograms** distort the shape of geographic regions so that the area directly encodes a data variable.

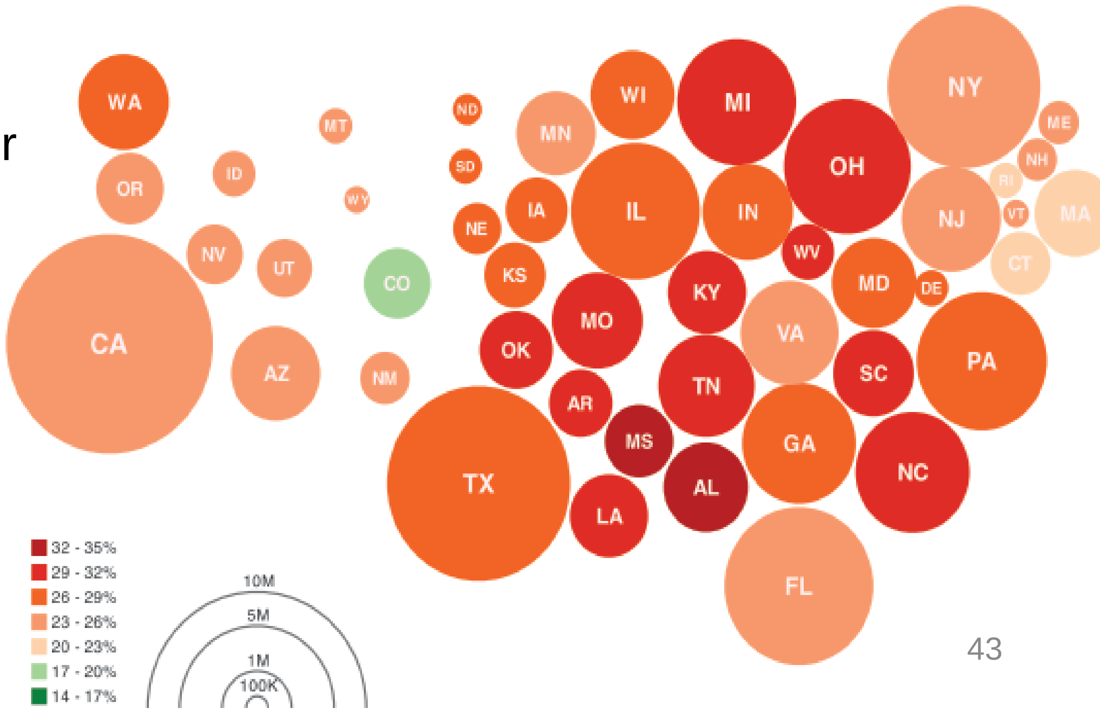
- There are several types

- **Dorling cartograms** represent each geographic region with a sized circle placed so as to resemble the true geographic configuration

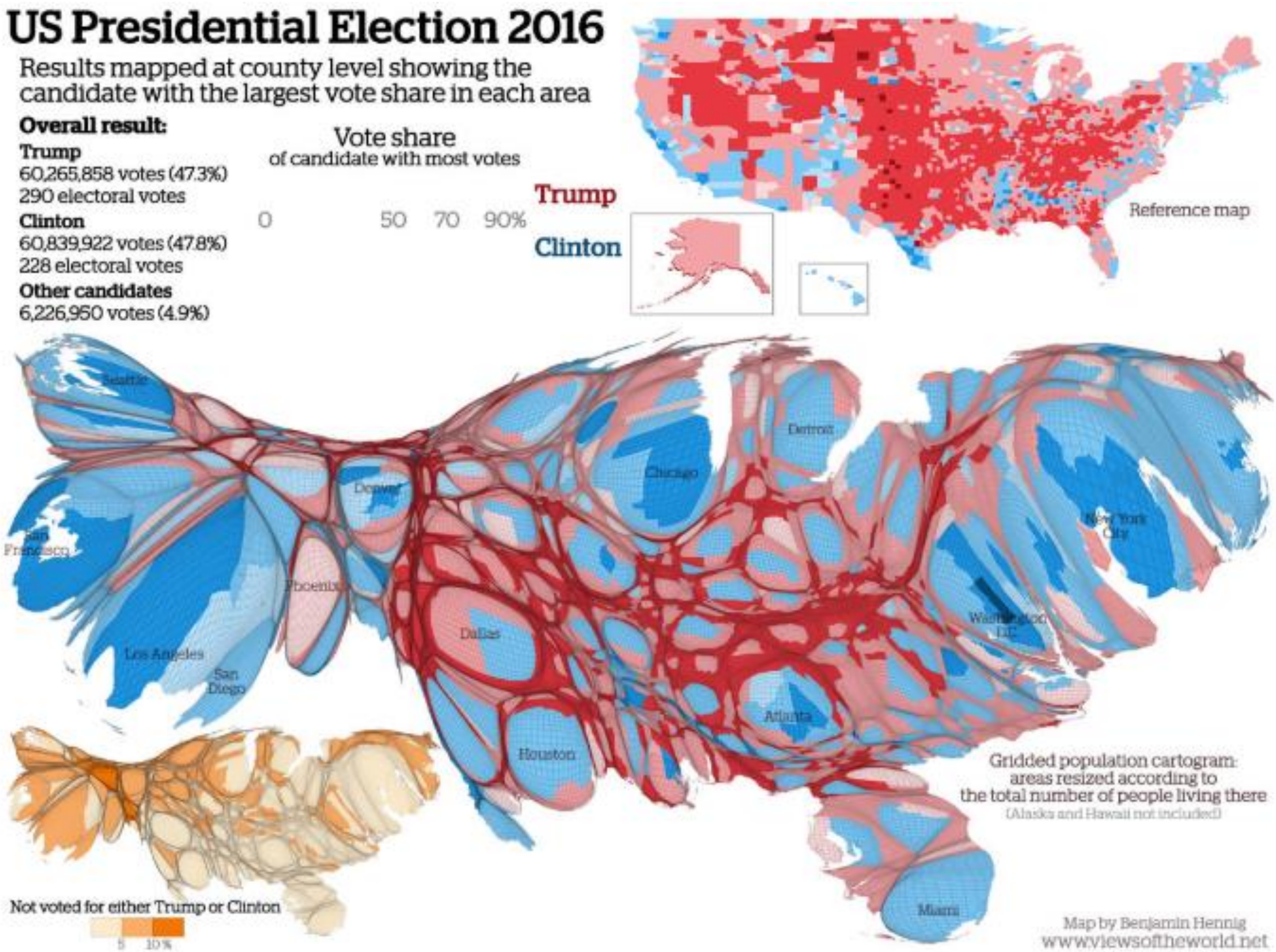
In these example:

- area encodes the total number of obese people per state

- color encodes percentage of obese population



Obesity in the US (2008)
(Heer et al., 2010)

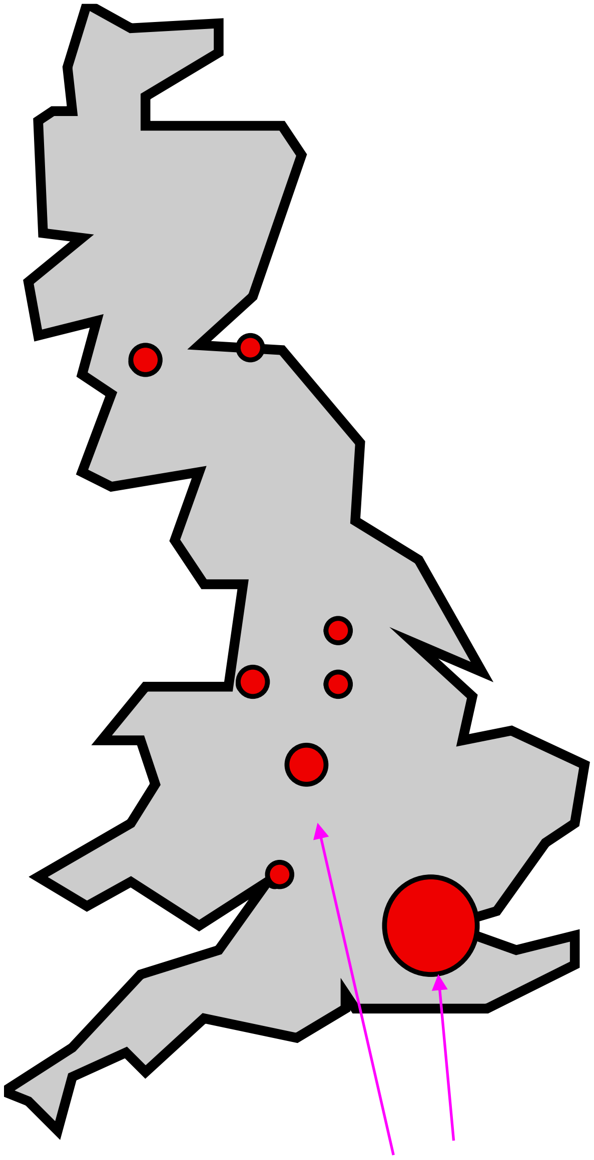


Cartogram showing the 2016 US election results (Click image for larger version)

https://geographical.co.uk/places/mapping/item/1981-us-election-cartogram-special

Some more examples on how humans see…

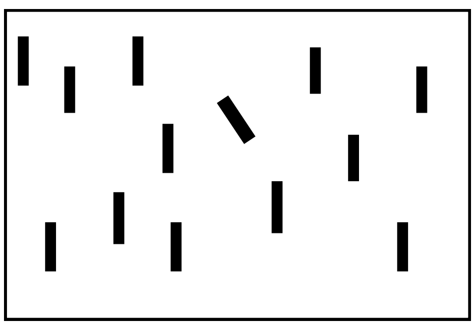
Population of major cities in England, Wales and Scotland. Circle area is proportional to population. (Spence, 2007)



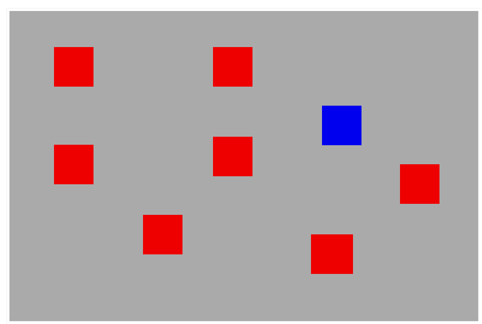
Things that “pop-out”

Pre-attentive processing: Things that "pop out"

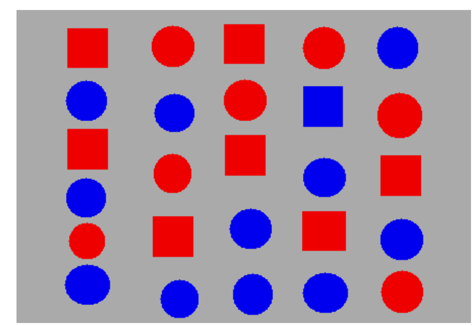
"We can do certain things to symbols to make it much more likely that they will be visually identified even after a very brief exposure" (Ware, 2004)

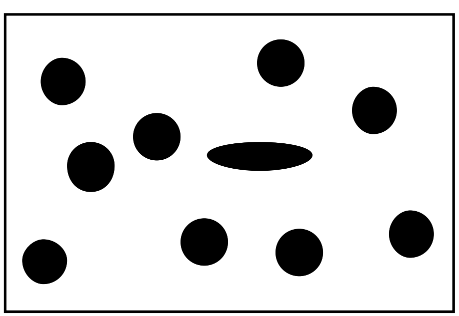


Orientation

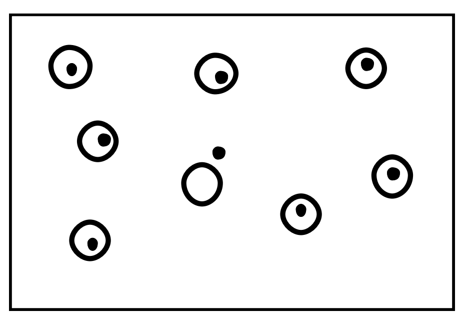


Colour





Shape



Enclosure

Where is the blue square?

(Spence, 2007)

But we should be careful…

Color is a strong visual cue

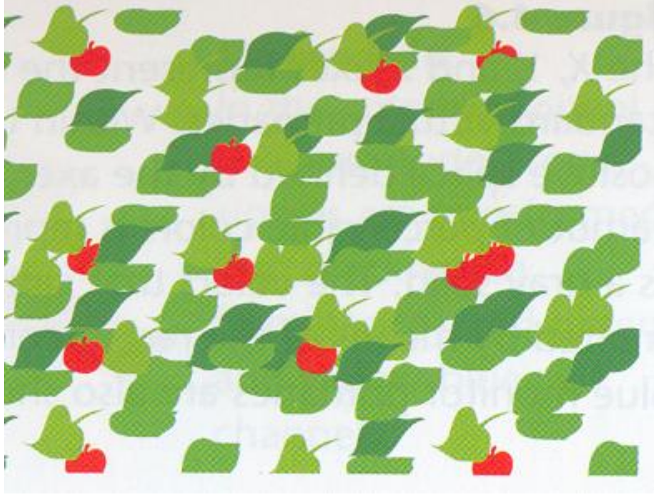
- How many cherries?

(Ware, 2004)

Color is a strong visual cue: it may help users perform their tasks

If correctly used

How many cherries?



Color may support users in many tasks!

Or not ...

Using color is complex as color perception is complex…

Color scales

Fundamental use cases for color in visualization:

- distinguish groups of data (qualitative color scales)
- represent data values (sequential color scales)
- Highlight (accent color scales)

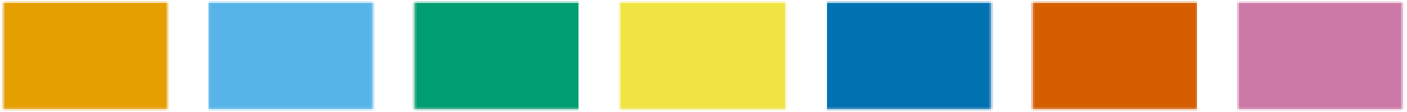
The types of colors and the way in which they should be used are quite different

(some examples next)

https://clauswilke.com/dataviz/

Qualitative color scales

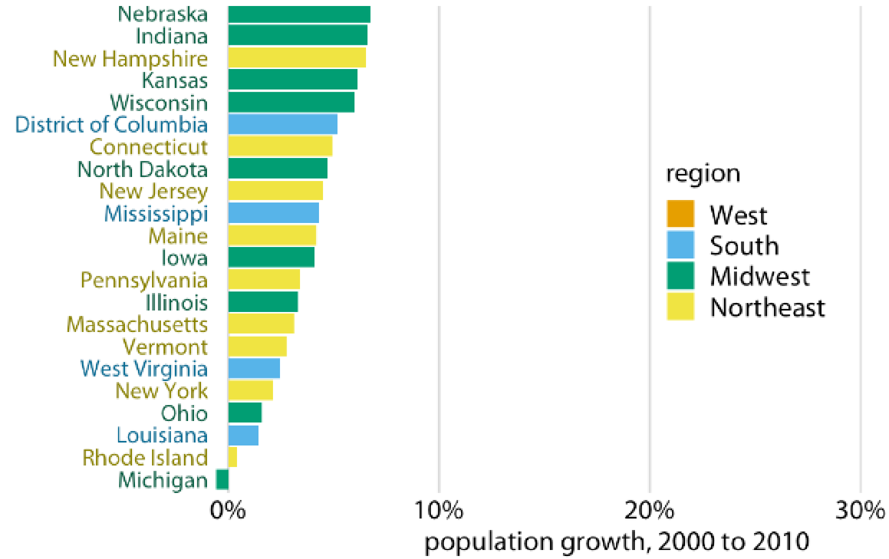
Okabe Ito



ColorBrewer Dark2

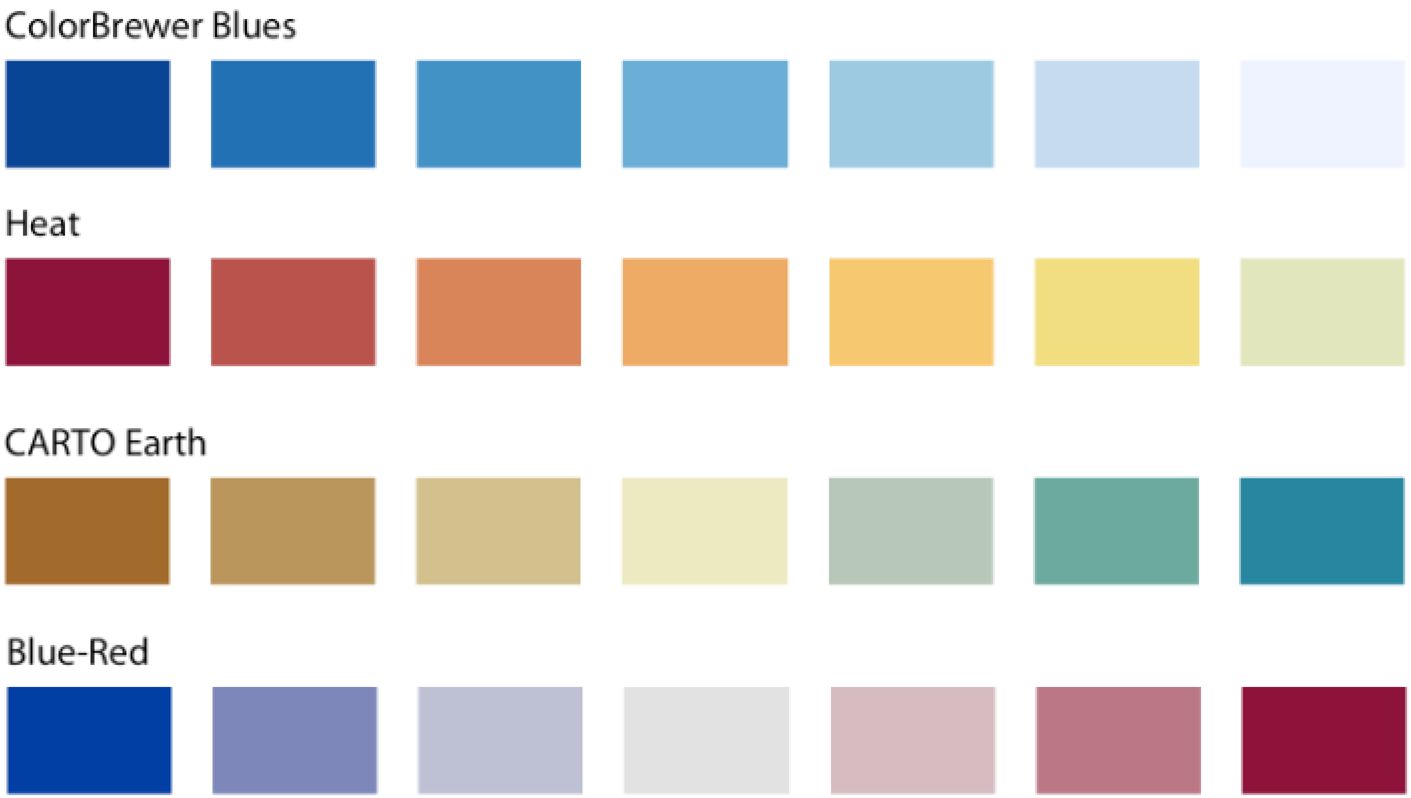
ggplot2 hue

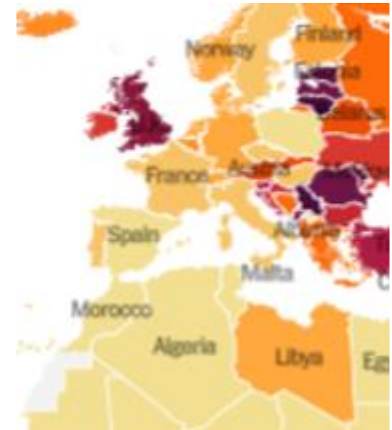
Colors are chosen to be clearly distinct and not stand out relative to others



https://clauswilke.com/dataviz/

Sequential color scales





monochrome

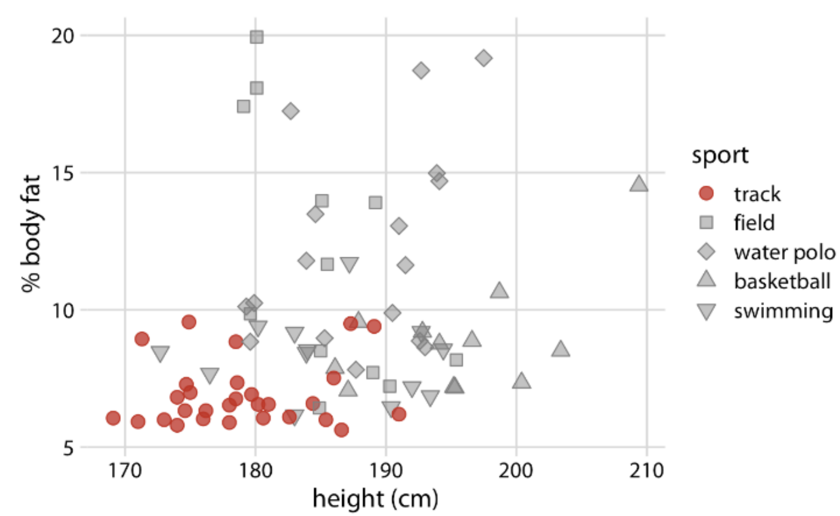
diverging

Colors should indicate which values are larger or smaller, and how distant two specific values are from each other, may be monochrome, diverging …

https://clauswilke.com/dataviz/

Accent color scales

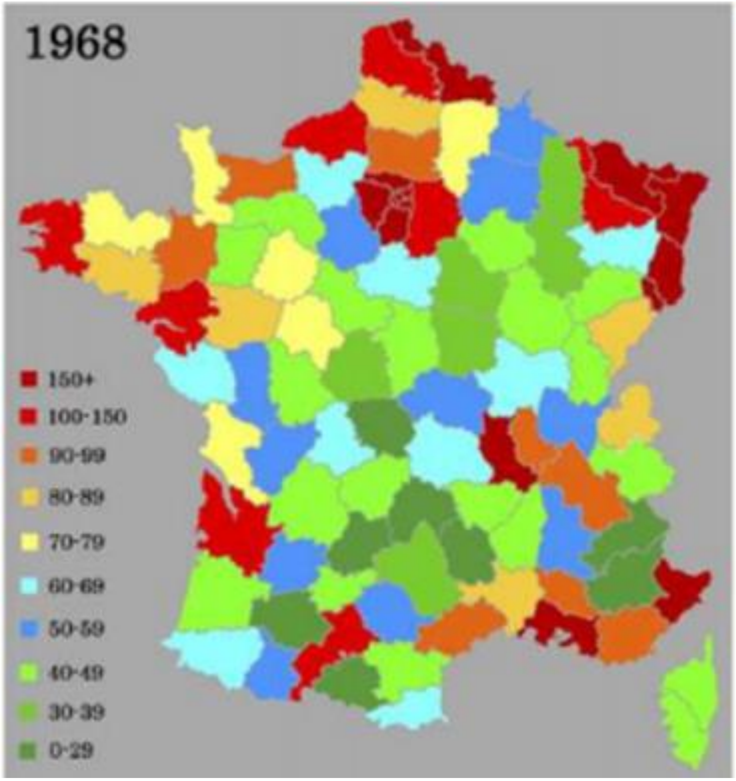
These scales contain a set of subdued colors and a matching set of stronger, darker, more saturated colors



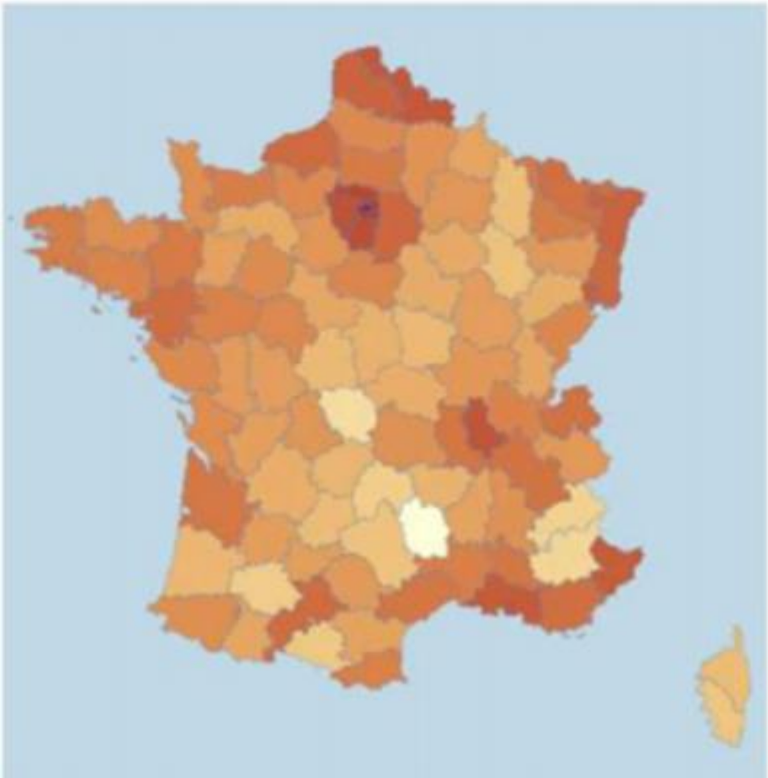
https://clauswilke.com/dataviz/

Color may not help or even make it more difficult!

A



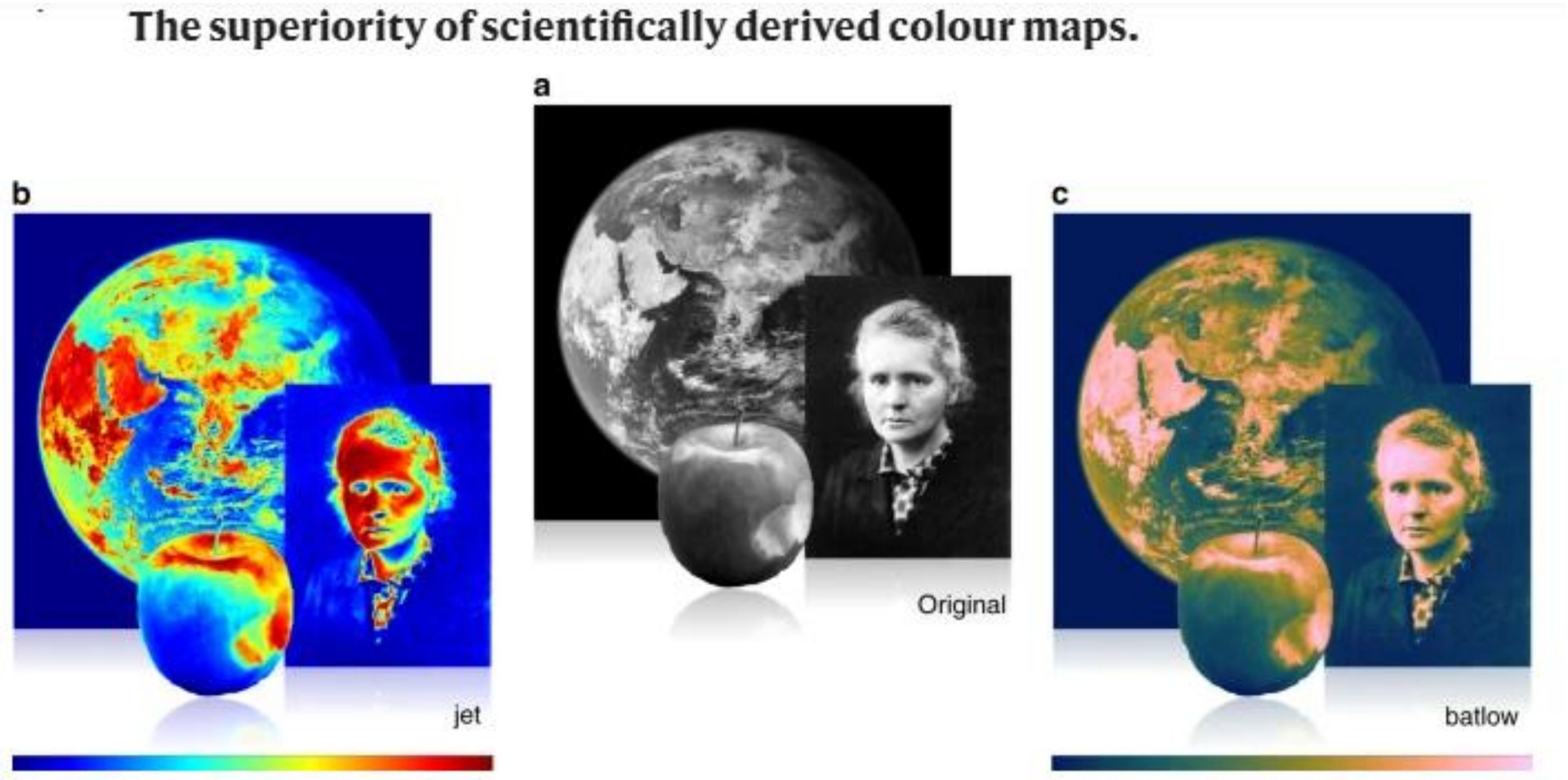
B



A- no preattentive association that allows efficiently determine the values (Kirk, 2012)

B- a single hue and a sequential color scheme representing values in an immediately understandable way

The misuse of colour in science communication



https://www.nature.com/articles/s41467-020-19160-7

Remember:

- Not everyone sees color:
- The most common form of color blindness is deuteranopia (“daltonism”)

- There are color blindness simulators:

  Try this one:

/

http://www.color-blindness.com/coblis-color-blindness-simulator

Normal vision

Deuteranopia

Tritanopia

http://www.colourblindawareness.org/

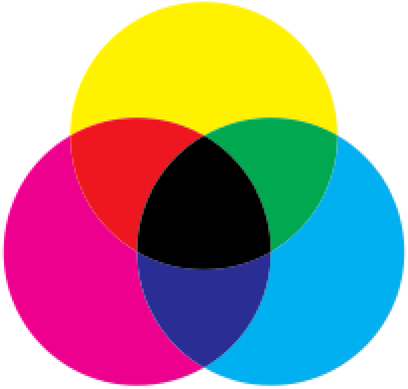
Some rules to use color in visualization

- Make it right is Black and White
- Less is more, or less is better
- Avoid using fully saturated colors in large numbers and in large areas
- Use fully saturated colors only when you want to highlight
- Use blue in larger areas and not in small areas
- Mind colorblindness and use simulators to test your designs
- When adopting color to distinguish, use colors that are easily distinct from each other
- …

Color models

Are used to measure and produce color

The basic H/W oriented models are RGB (additive) and CMY (subtractive)



RGB: emitting/screens/projectors CMYK: reflecting/printers

Are not related to human perception, but to the physical process

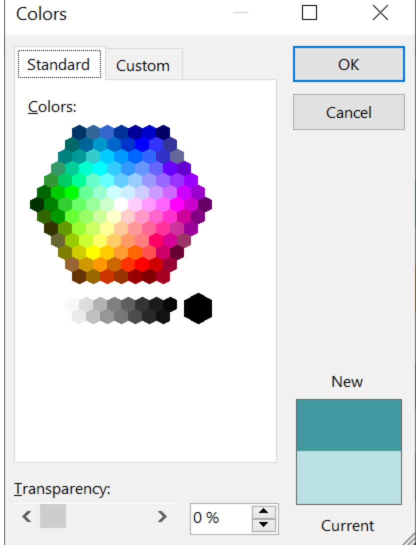
Should not be used directly to produce color scales

There are color models (HSV and HSL) based on perceptual variables:

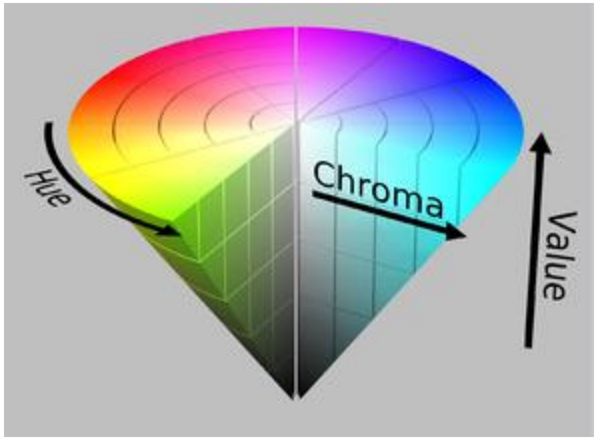
hue (violet, blue, green, yellow, red …)

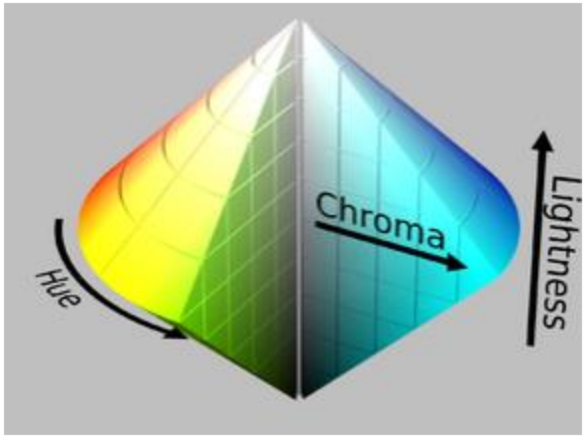
saturation (amount of white)

value/brightness



used when we intuitively describe colors (e.g. light blue or dark green)





HSV

HLS

Are more adequate for users to specify color

Other models (perceptually corrected) are better to specify quantitative color scales (e.g. L*, a*, b* color model)

Representing Hypervariate (or multivariate) data

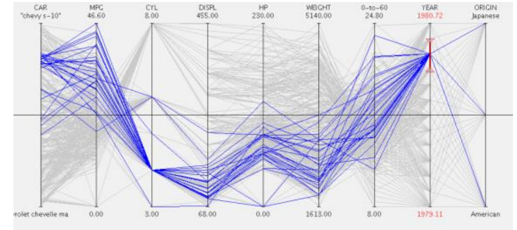
- Many real problems are of high dimensionality
  (even after reducing dimensionality…)

- The challenge of representing hypervariate data is substantial and continues to stimulate invention

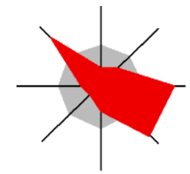
- Some of the mentioned representation techniques can be scaled to represent hypervariate data (to a limited extent)

Techniques for Hypervariate (or multivariate) data Visualization

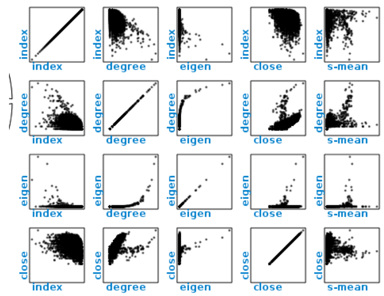
- Coordinate plots — parallel coordinate plots



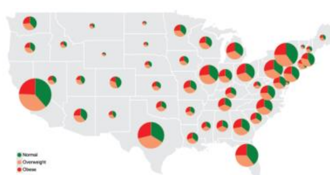
star (radar, or spider) plots



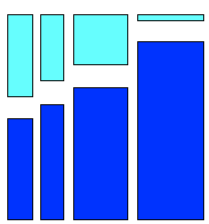
- Scatterplot Matrix



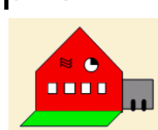
- Maps

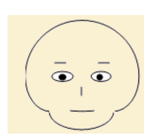


- Mosaic Plots



- Icons/glyphs





Representing Hypervariate

- **Consider dimensionality reduction!**

- Several methods can be used (e.g.):

  - Principal Components Analysis

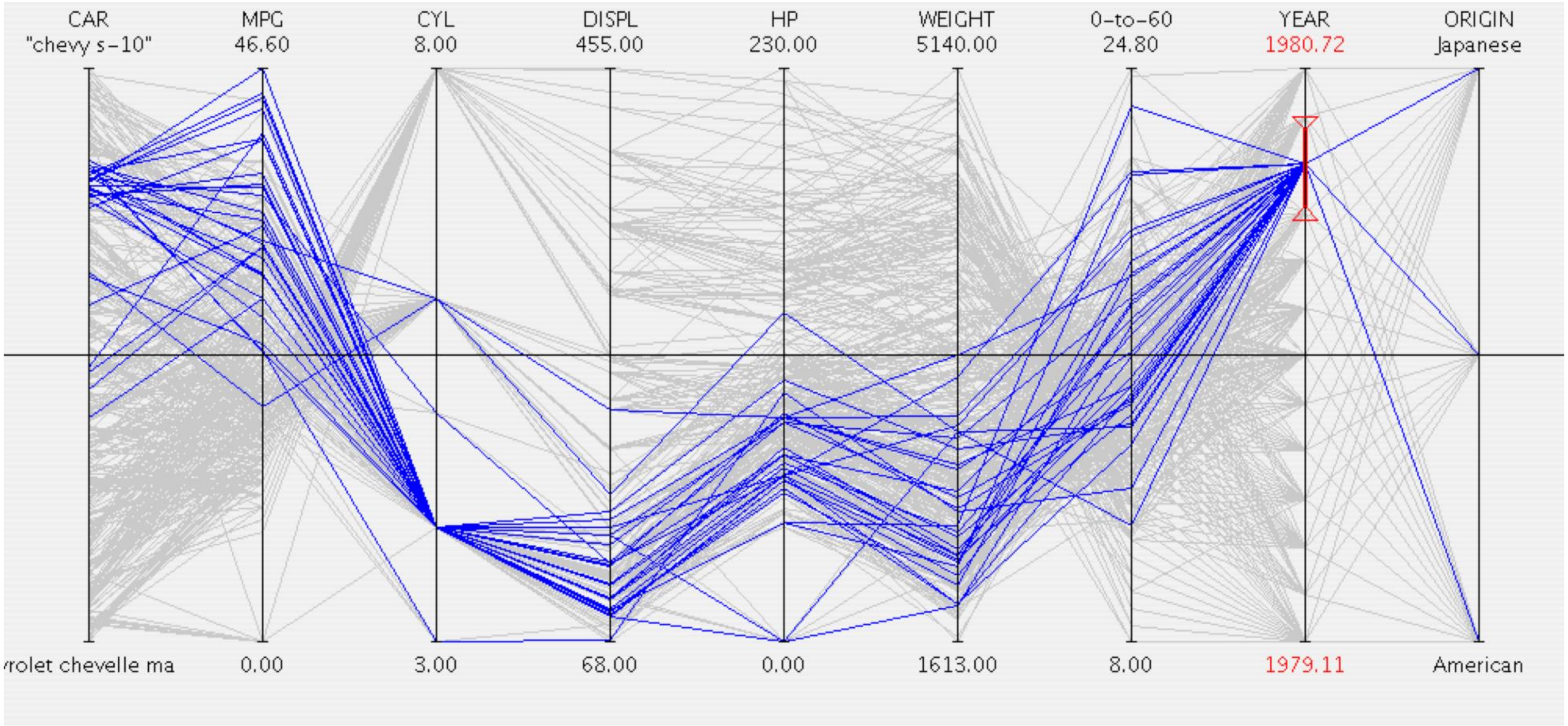
  - t-SNE (t-distributed stochastic neighbor embedding)

  …

- Parallel coordinates plots are one of the most popular techniques for hypervariate data

- They have a very simple basis

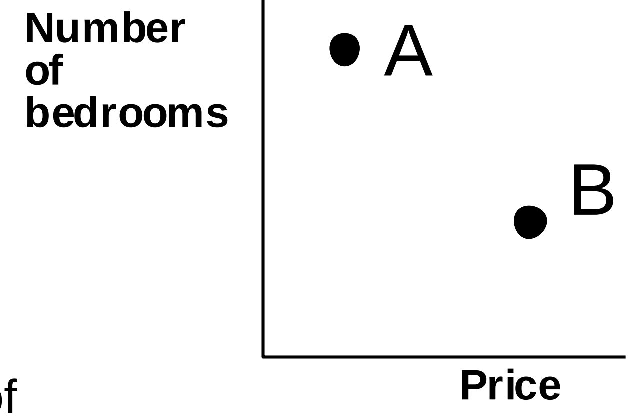
Make	Price (£)	MPG	Rating	Age (yrs)	…
Ford	*15,450*	*31*	*****	*3*	
Chevy	*12,450*	*27*	***	*4*	
…					



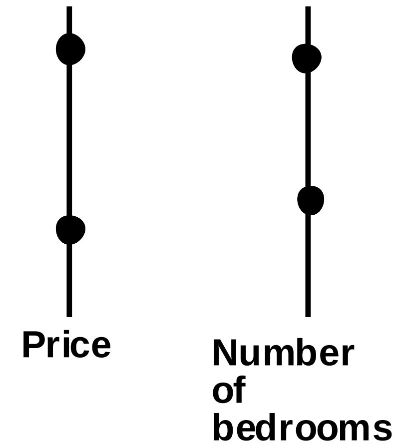
(Spence, 2007)

Consider a simple case of bivariate data:

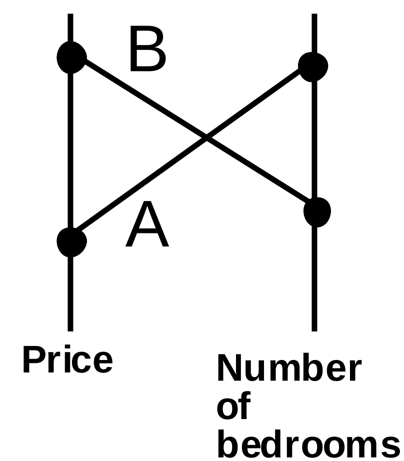
1- A scatterplot represents the price and number of bedrooms associated with two houses



2- the axes are detached and made parallel; each house is represented by a point on each axis

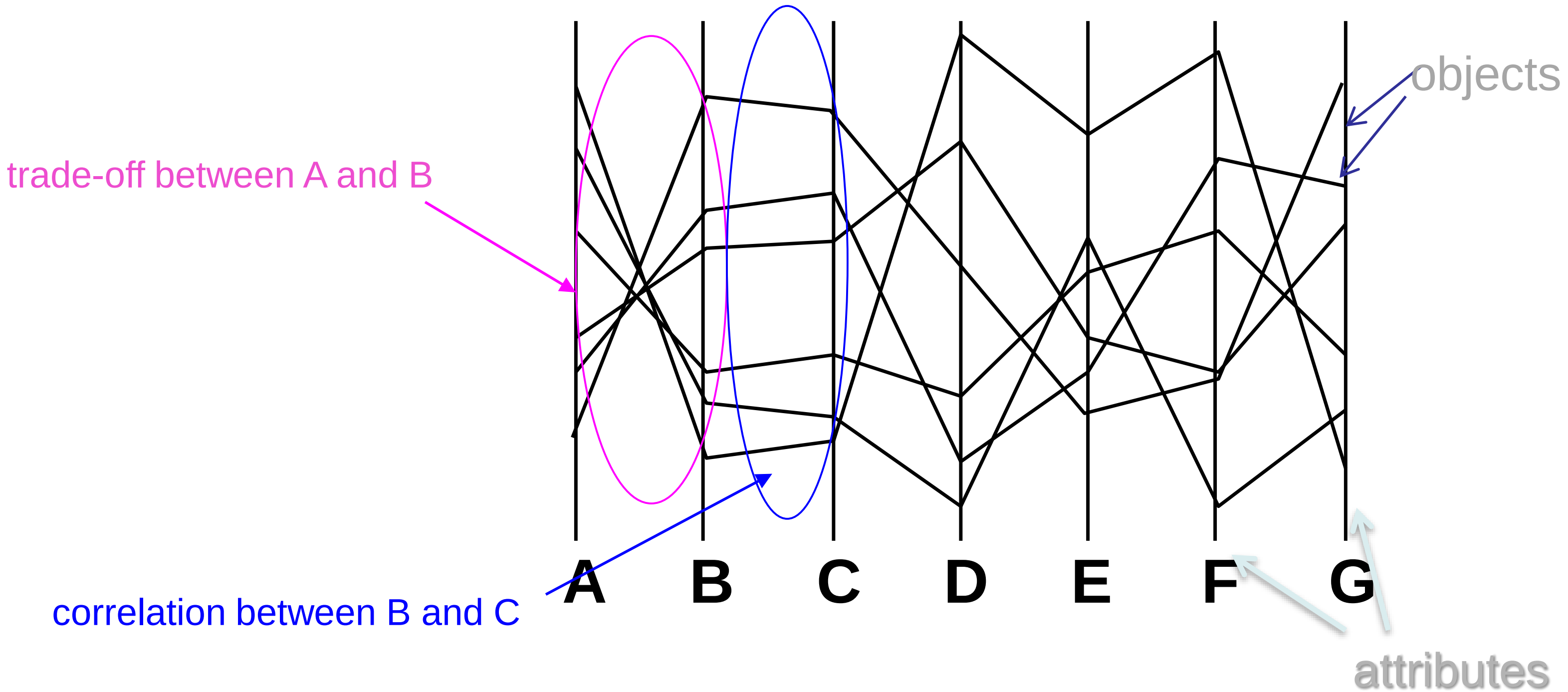


3- To avoid ambiguity the pair of points representing a house are joined and labeled

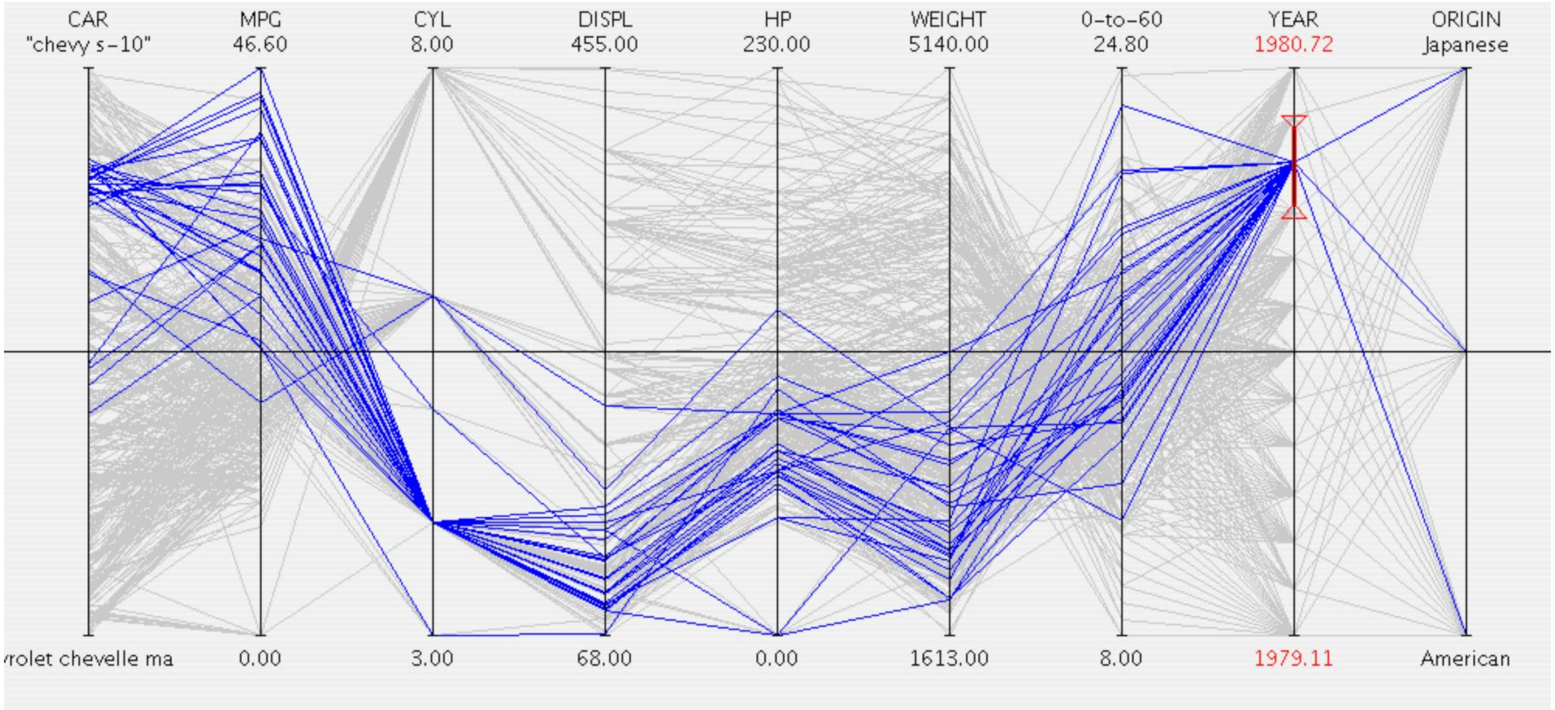


- For objects characterized by many attributes the parallel coordinate plots offer many advantages

A example for six objects, each characterized by seven attributes:

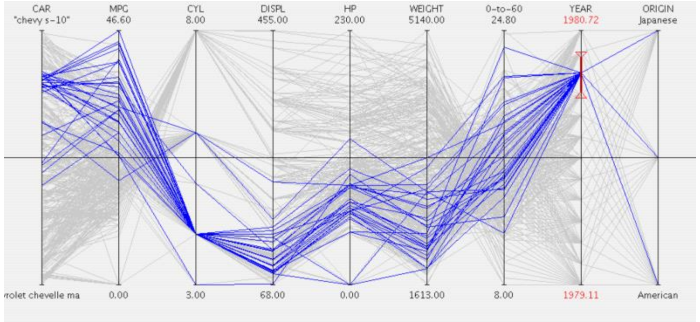


The trade-off between A and B, and the correlation between B and C, are immediately apparent. The trade-off between B and E, and the correlation between C and G, are not.



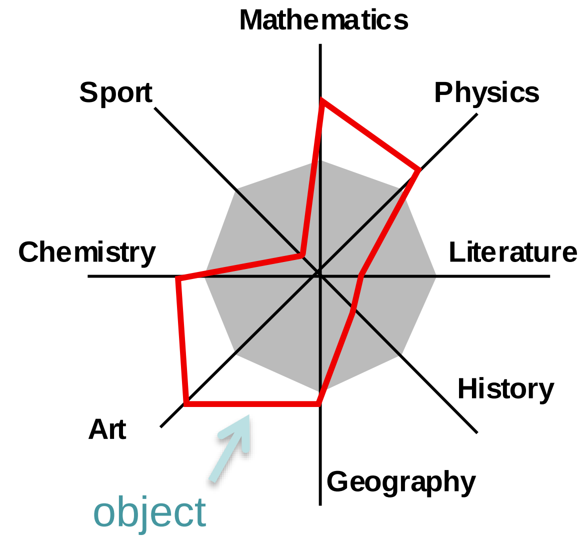
A parallel coordinate plot representation of a collection of cars, in which a range of the attribute *Year* has been selected to cause all those cars manufactured during that period to be highlighted.

Properties of parallel coordinate plots:



- Suitable to identify relations between attributes
- Objects are not easily discriminable; each object is represented by a polyline which intersects many others
- They offer attribute visibility (the characteristics of the separate attributes are particularly visible)
- The complexity of parallel coordinate plots (number of axes) is directly proportional to the number of attributes
- All attributes receive uniform treatment

- Star plots have many features in common with parallel coordinate plots



(Spence, 2007)

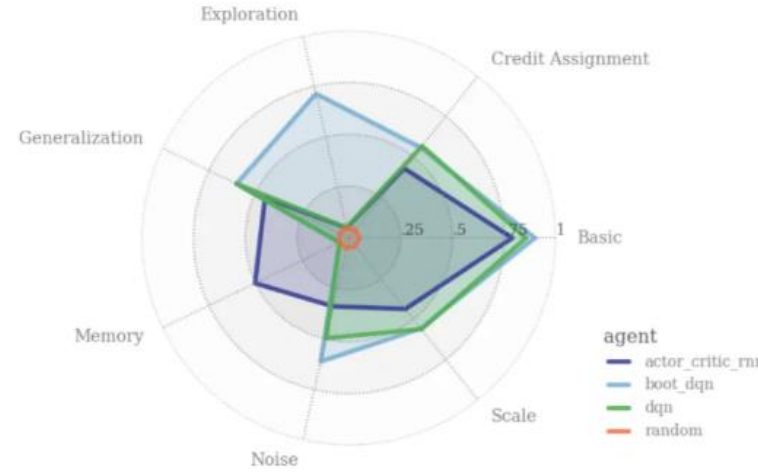
- An attribute value is represented by a point on a coordinate axis

- Attribute axes radiate from a common origin

- For a given object, points are joined by straight lines

- Other useful information such as average values or thresholds can be encoded

Properties of star plots:

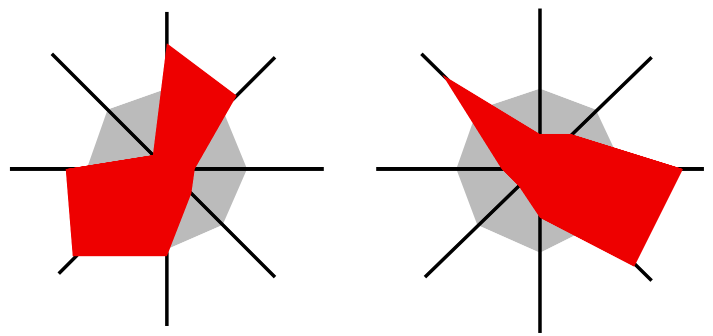


https://syncedreview.com/2019/08/16/deepmind-bsuite-evaluates-reinforcement-learning-agents/

- Their shape can provide a reasonably rapid appreciation of the attributes of the objects

- They offer object visibility and are suitable to compare objects

(by visibility it is meant the ability to gain insight pre-attentively; without a great cognitive effort)

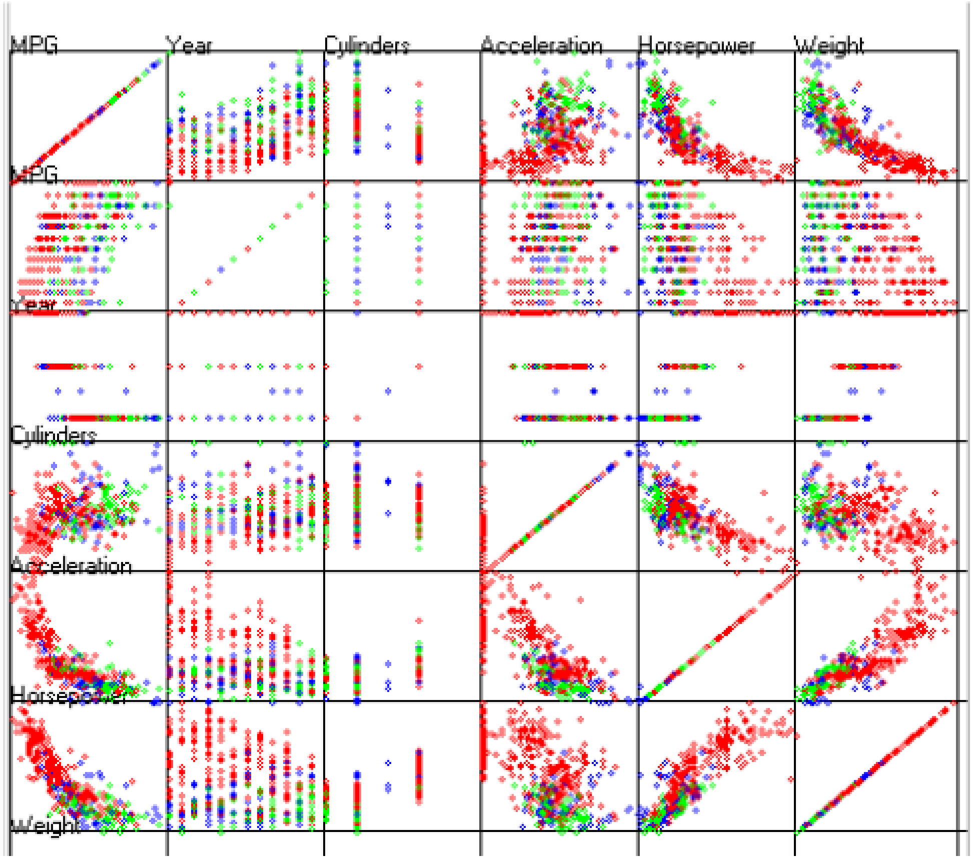


Bob's performance Tony's performance

(Spence, 2007)

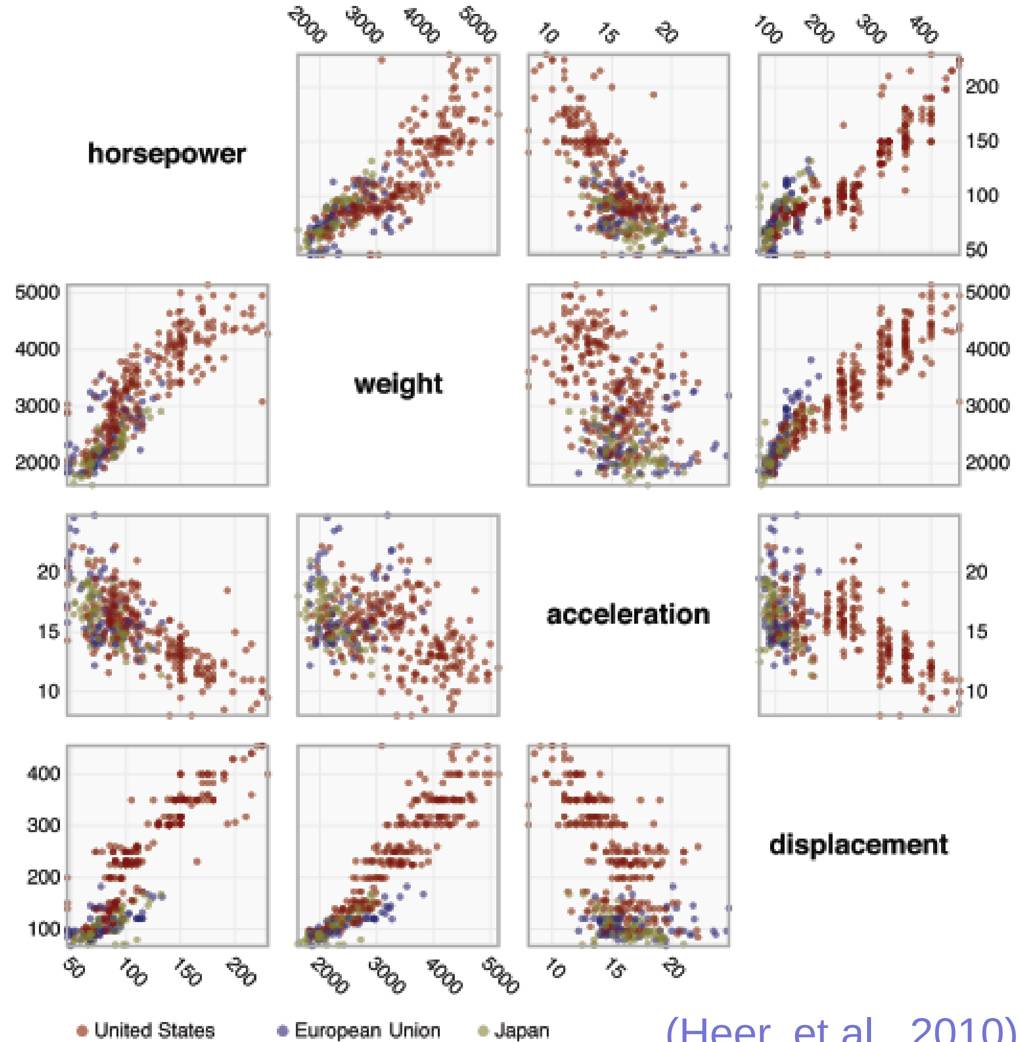
- The scatterplot matrix (SPLOM) is applicable to higher dimensions
- However, as the number of attributes increase, the number of different pairs of attributes increases rapidly:
  - 2 attributes -> 1 scatterplot
  - 3 attributes -> 3 scatterplots
  - 4 attributes -> 6 scatterplots

We may try to reduce the number of dimensions keeping the more relevant



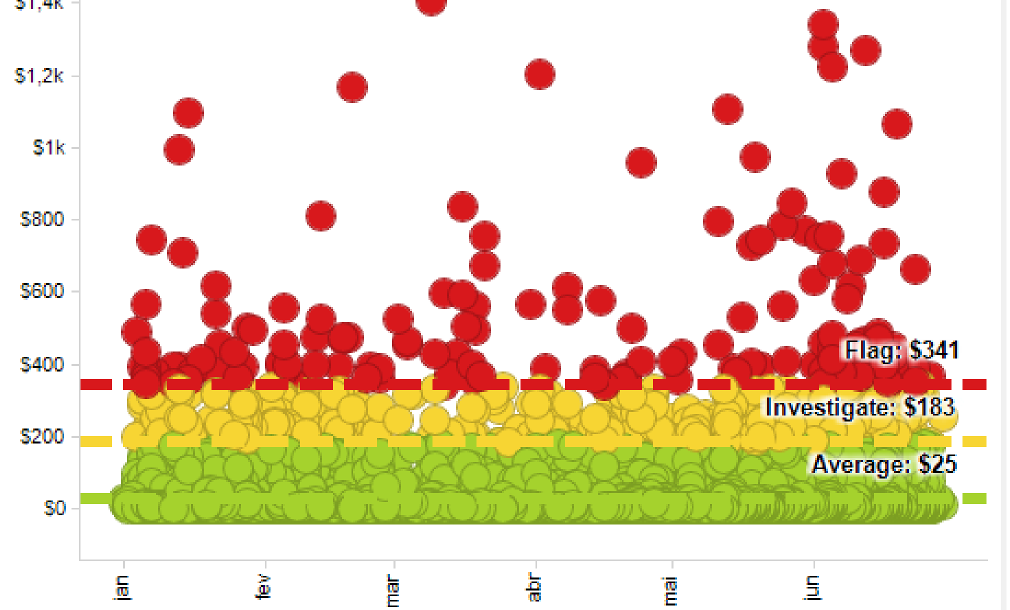
Scatterplot matrix for 6 attributes of a car dataset

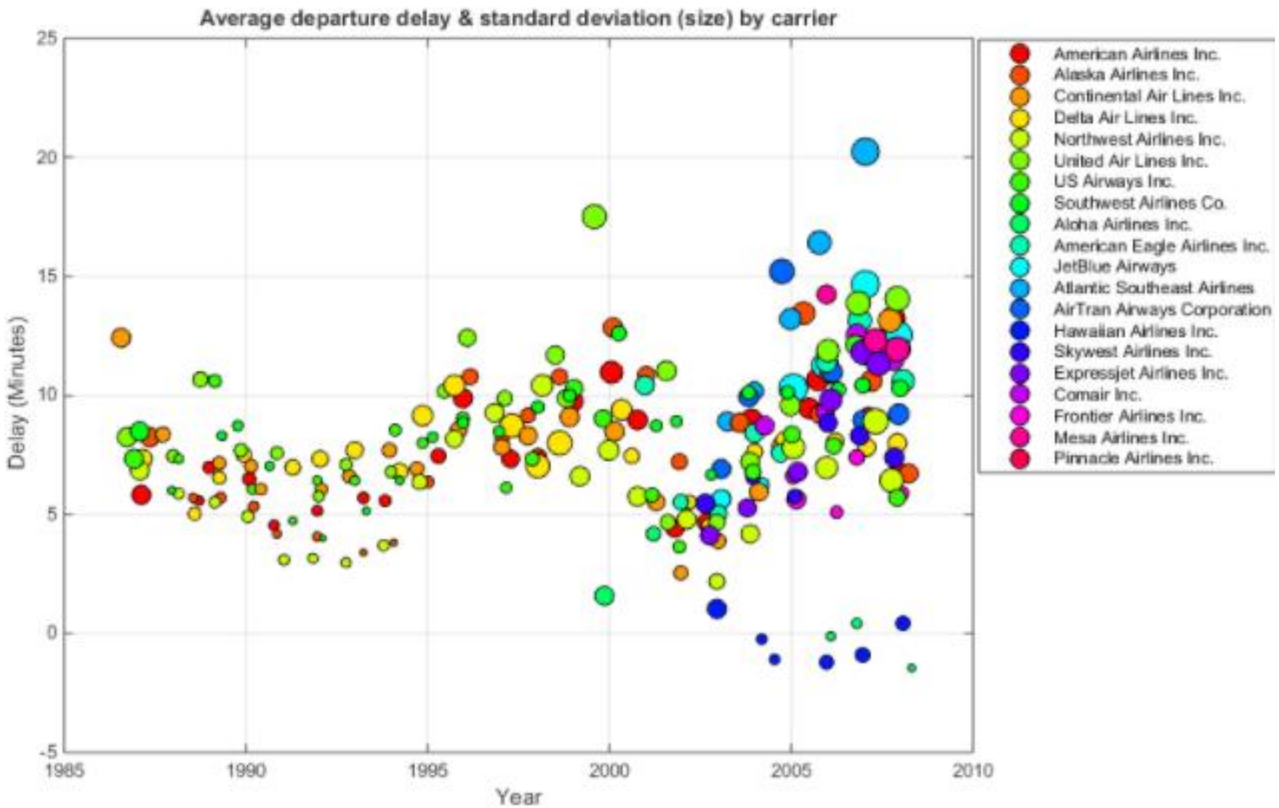
- Another example of Scatterplot matrix for a car dataset



(Heer, et al., 2010)

- A single scatterplot can be used together with other encoding techniques to represent data of higher dimension





https://www.mathworks.com/matlabcentral/fileexchange/48005-bubbleplot-multidimensional-scatter-plots

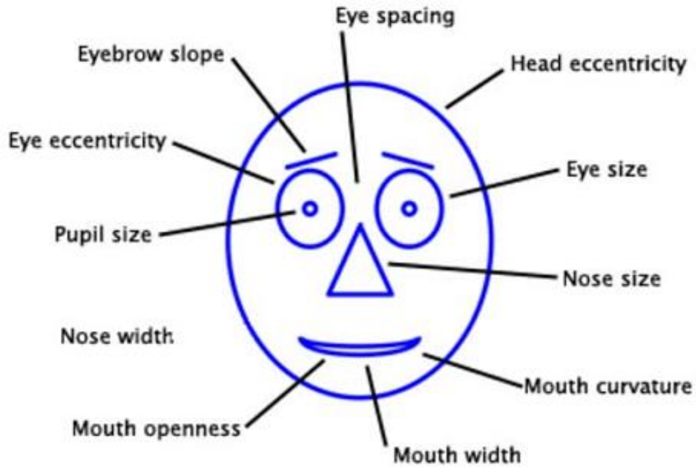
A scatterplot representing 5 variables

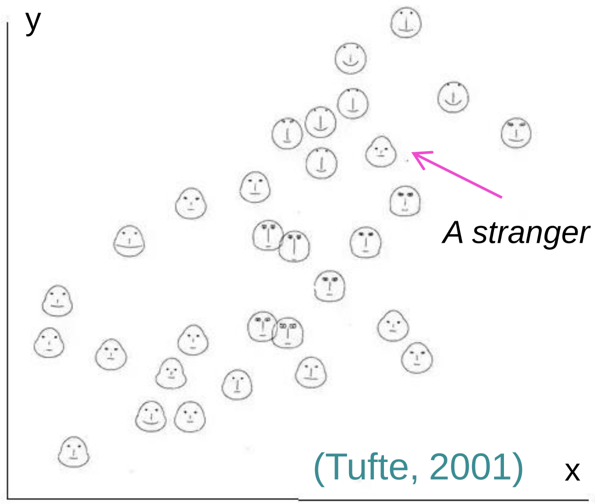
Hans Rosling's 200 Countries, 200 Years, 4 Minutes: 120 000 values

Income (x), Age expectancy (y) , Time (t), Continent (color), Population (size of circle)

https://www.youtube.com/watch?v=jbkSRLYSojo

Icons (aka glyphs) represent a number of attributes qualitatively or quantitatively



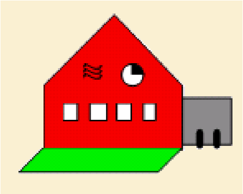


Chernoff Faces allow attribute values to be encoded in the features of cartoon faces

They were originally used to study geological samples, each characterized by 18 attributes

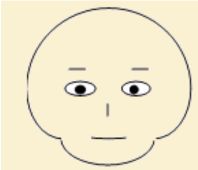
(https://en.wikipedia.org/wiki/Chernoff_face)

- Two examples of metaphorical icons:

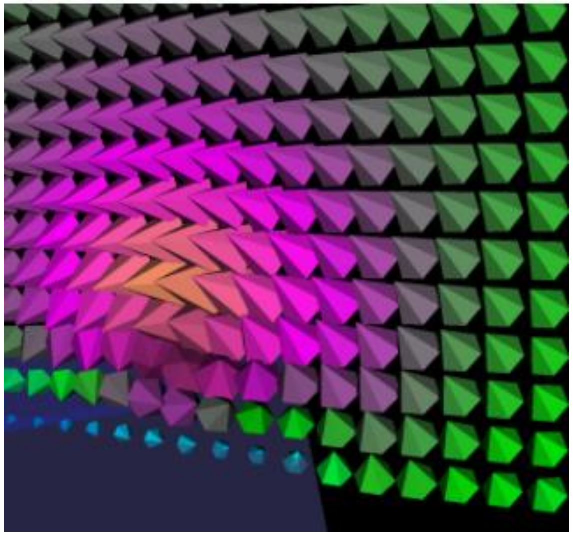


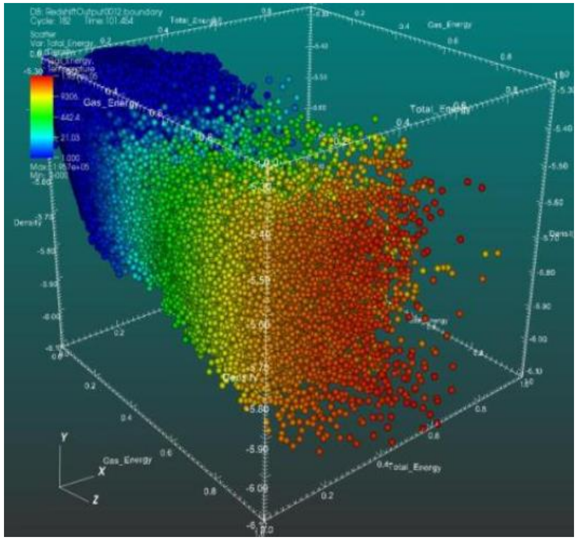
  - with direct relation between icon and object (house icon)

  - no direct relation between facial features and attributes they represent (Chernoff faces)



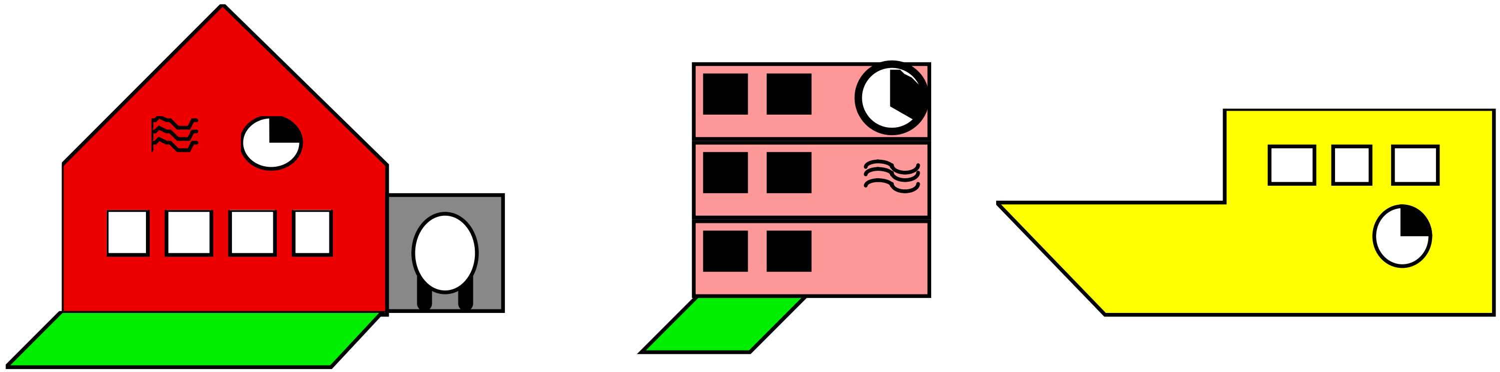
- Examples in SciVis





https://en.wikipedia.org/wiki/Glyph_(data_visualization)

Multidimensional icons representing eight attributes of a dwelling



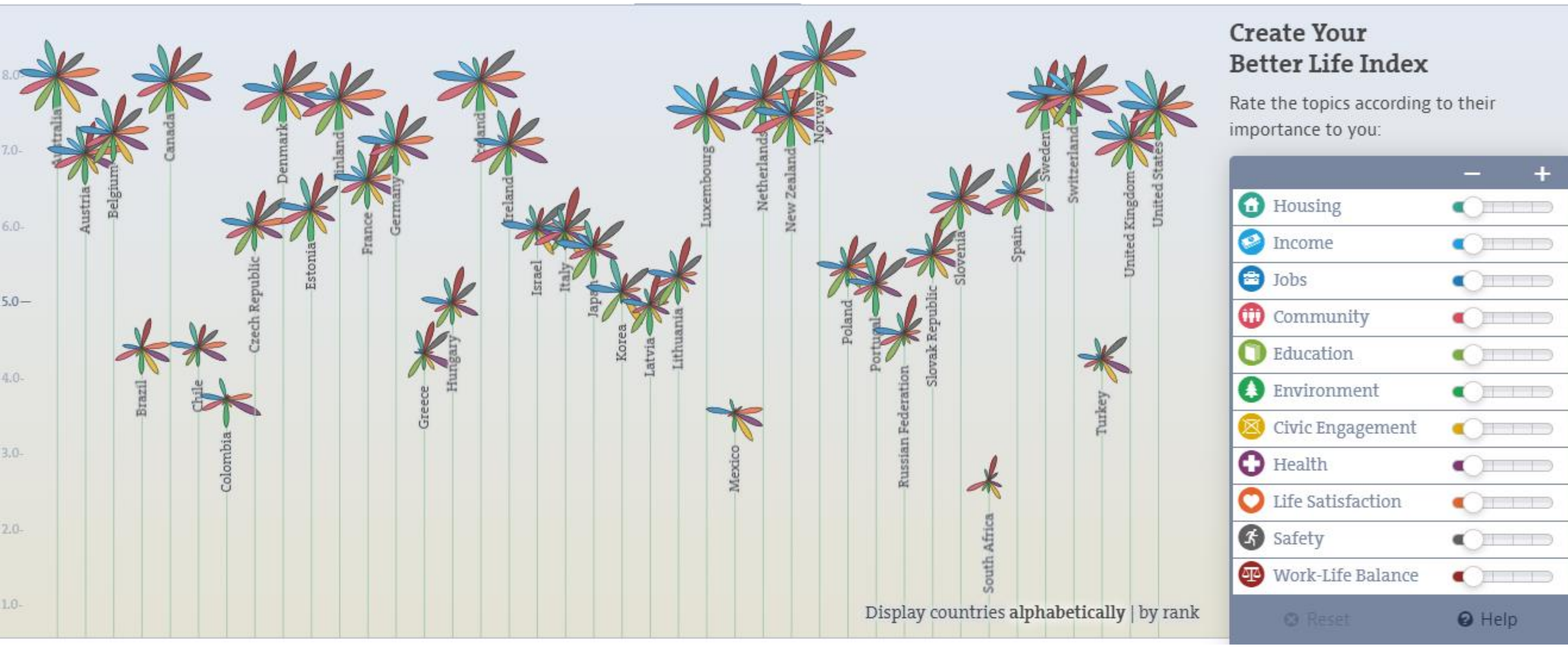
house	flat	houseboat
£400,000	£300,000	£200,000
garage	no garage	no garage
central heating	central heating	no central heating
four bedrooms	two bedrooms	three bedrooms
good repair	poor repair	good repair
large garden	small garden	no garden
Victoria 15 mins	Victoria 20 mins	Victoria 15 mins

Textual descriptions of the dwellings represented by the multidimensional icons (Spence, 2007)

Glyph chart example:

Based on a shape being the main artifact of representation

The physical properties of the shape represent different categorical variables sized according to the associated quantitative value and distinguished through color



http://oecdbetterlifeindex.org

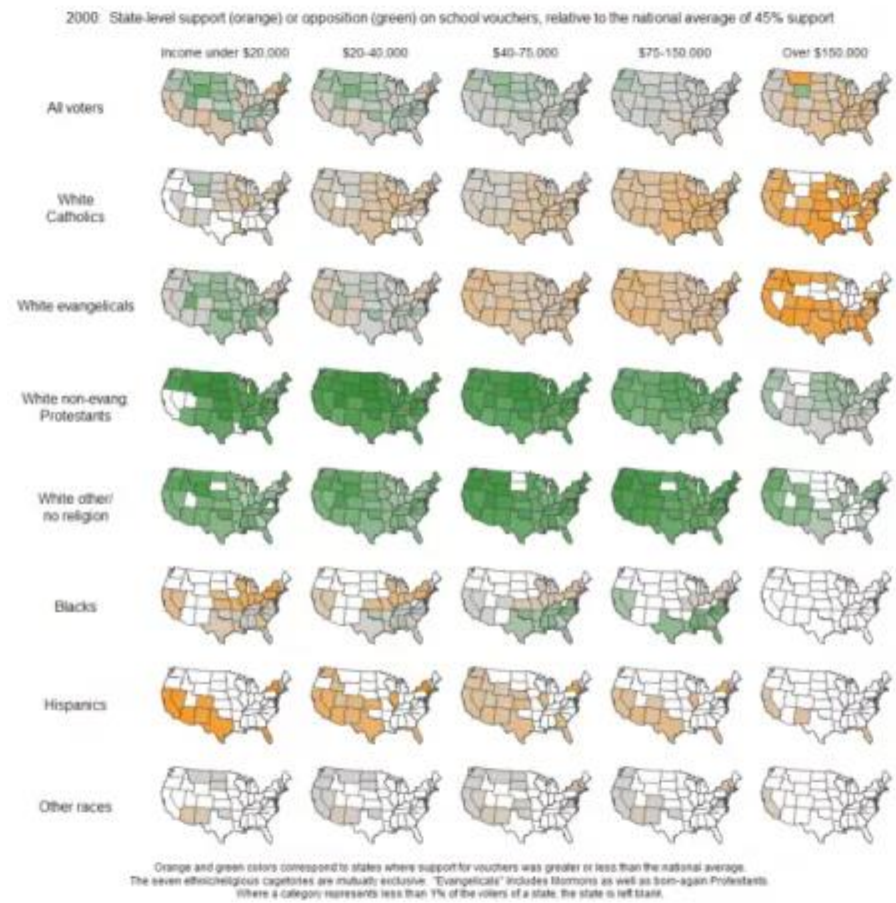
(Kirk, 2012)

Useful arrangement of several charts

Small multiples:

arrangement approach that facilitates efficient and effective comparisons

(Kirk, 2012)



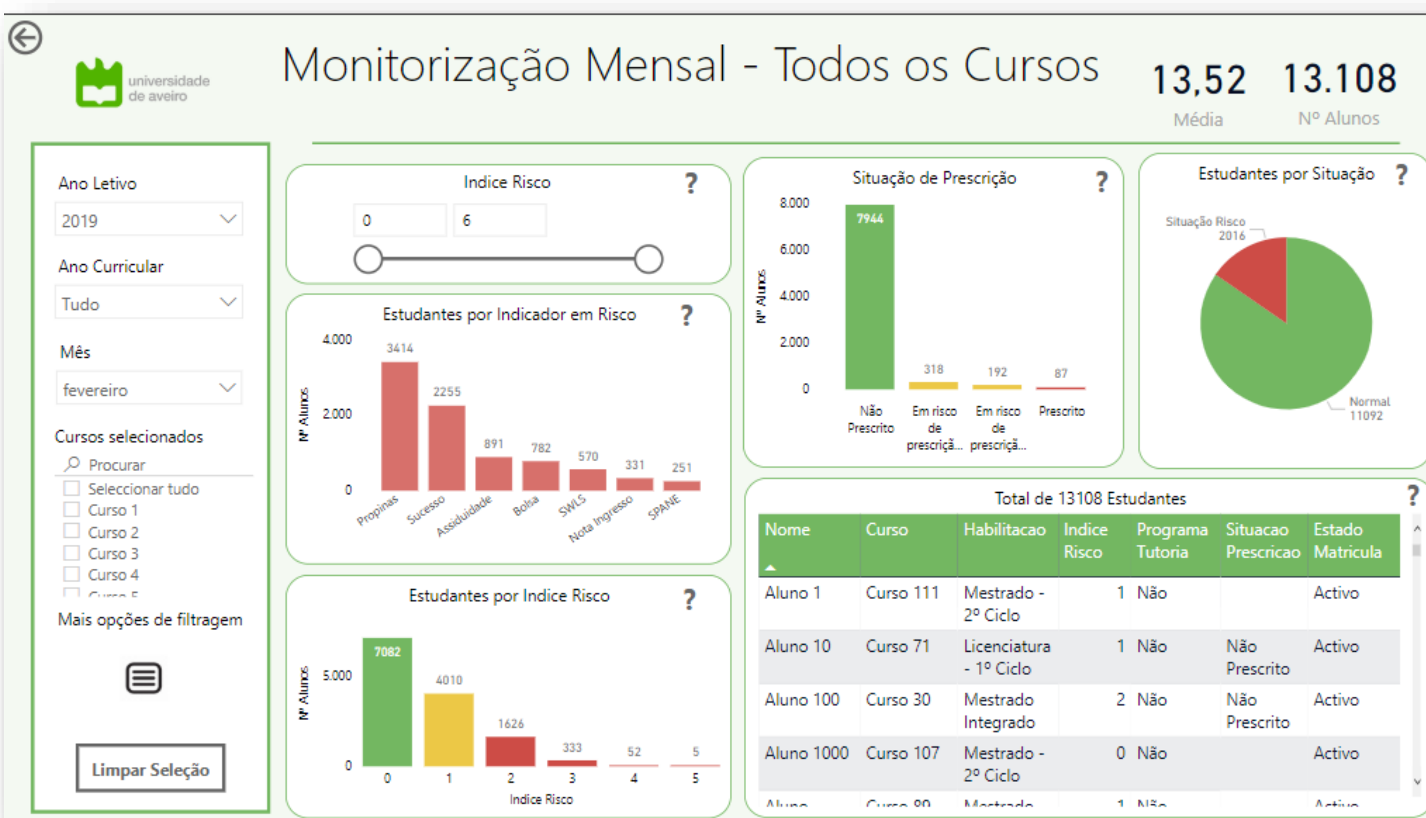
Dashboards

Visual display summarizing a dataset providing information at-a-glance (e.g. KPIs)

" A dashboard is a visual display of the most important information needed to achieve one or more objectives; consolidated and arranged on a single screen so the information can be monitored at a glance. " (Few, 2004)

https://www.nngroup.com/articles/dashboards-preattentive/

Prototype: “Portal dos indicadores, UA”



As seen by people with green-blind deuteranopia; does it work?

https://www.color-blindness.com/coblis-color-blindness-simulator/

Example:
Use visualization techniques to help answer the following questions:

Is there a relation between wanted salary and experience?
How many candidates ask for a salary in [30000, 50000] and in [55000, 75000]?
How many candidates have an advanced level of English?

	Education	**Age**	**Prof. Experience**	**English**	**Wanted salary**
#	**(MSc/PhD)**	**(years)**	**(years)**	**(Bas/Adv)**	**($$/year)**
1	MSc	22	0	Advanced	36000
2	MSc	23	0	Basic	36000
3	MSc	24	1	Advanced	36000
4	PhD	30	7	Advanced	72000
5	MSc	25	1	Basic	40000
6	PhD	29	5	Advanced	60000
7	MSc	31	7	Advanced	55000
8	MSc	23	0	Advanced	36000
9	MSc	26	2	Intermediate	40000
10	PhD	32	9	Intermediate	65000
11	BSc	30	7	Intermediate	30000
12	PhD	40	17	Advanced	80000
13	MSc	28	4	Advanced	40000

the complete table has many more candidates and attributes, but you may test with these

Bibliography

- Kirk, A., *Data Visualization: a successful design process*, Packt Publishing, 2012
- Kirk, A., *Data Visualization: a handbook for data driven design* * , Sage, 2016
- Mazza, R., *Introduction to Information Visualization*, Springer, 2009
- Munzner, T., *Visualization Analysis and Design* * , A K Peters/CRC Press, 2014
- Spence, R., *Information Visualization, Design for Interaction*, 2nd ed., Prentice Hall, 2007
- Spence, R., *Information Visualization, an Introduction*, 3rd ed., Springer, 2014
- Ware, C., *Information Visualization* * , 3rd ed., Morgan Kaufmann, 2013
- Wilke, C., *Fundamental of Data Visualization*, 2019
- Tufte, E., *The Visual Display of Quantitative Data*, 2nd ed, Graphics Press, 2001

- Books with * and other interesting books at:

https://learning.oreilly.com/playlists/74bfec5e-4346-48ff-82b4-657fda6922b6