Computer Aided Archaeology

05 - Visualisation I

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

27/10/21

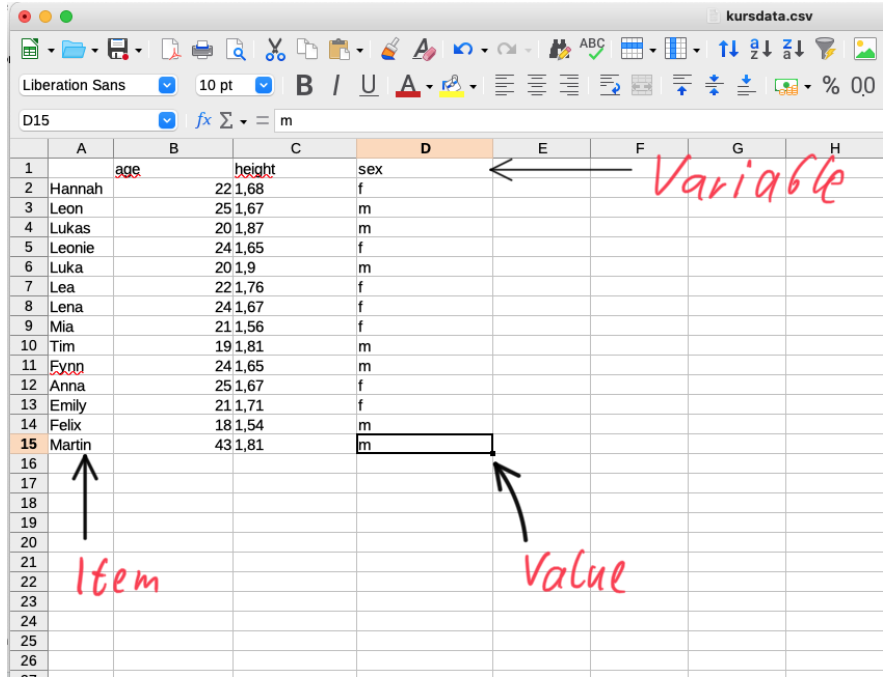
Why data visualisation

Converting raw data to a form that is viewable and understandable to humans

- humans are visual animals -> we evolved to identify patterns visually
- helps to map complex information into an meaningful picture
- enables to "see" more than two dimensions of the data and their interplay

Data, variables, values

- variable:
  - What ist measured or analysed.
  - e.g. height
- item:
  - That whichs variable is measure
  - e.g. me as „possessor“ of a height, graves, persons...
- values:
  - The actual measurement.
  - e.g. my height is 1.81 m.



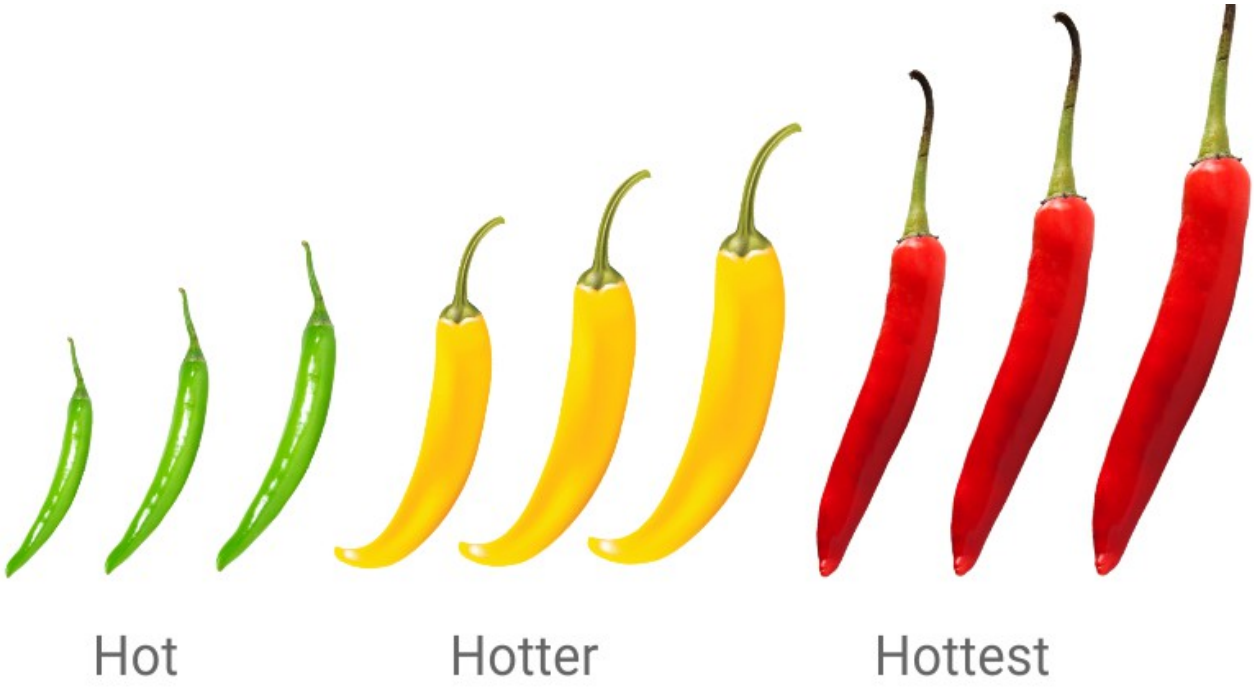
	A	B	C	D	E	F	G	H
1		age	height	sex				
2	Hannah	22	1,68	f				
3	Leon	25	1,67	m				
4	Lukas	20	1,87	m				
5	Leonie	24	1,65	f				
6	Luka	20	1,9	m				
7	Lea	22	1,76	f				
8	Lena	24	1,67	f				
9	Mia	21	1,56	f				
10	Tim	19	1,81	m				
11	Fynn	24	1,65	m				
12	Anna	25	1,67	f				
13	Emily	21	1,71	f				
14	Felix	18	1,54	m				
15	Martin	43	1,81	m				

Levels of measurement

nominal or categorical:

- You can only decide if something belongs to a category
- Categories which do not have a defined relationship among each other, only counting is possible (e.g. sex)

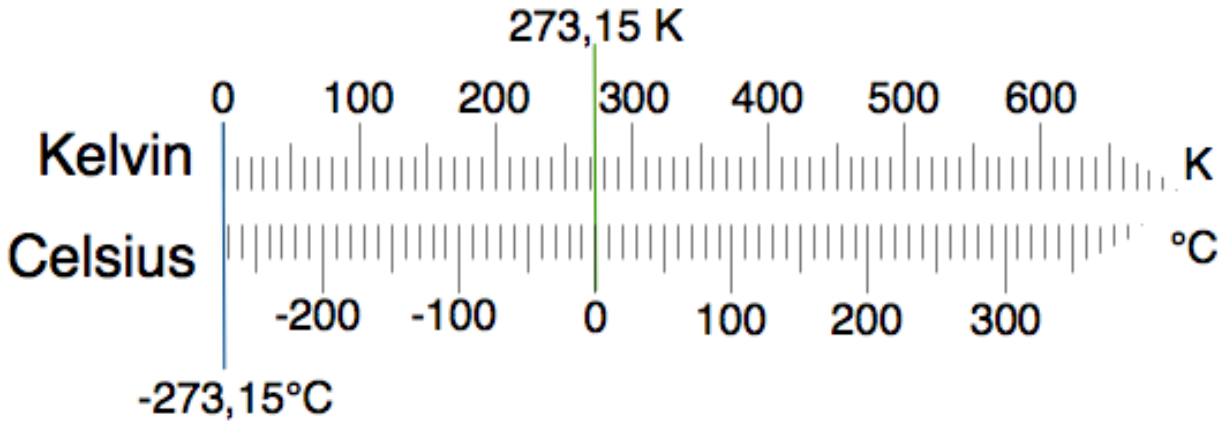
Levels of measurement



ordinal:

- Categories which are comparable and differ from each other in their characteristic [size/power/intensity]
- their rank is determinable (e.g. preservation conditions – bad < medium < good)

Levels of measurement



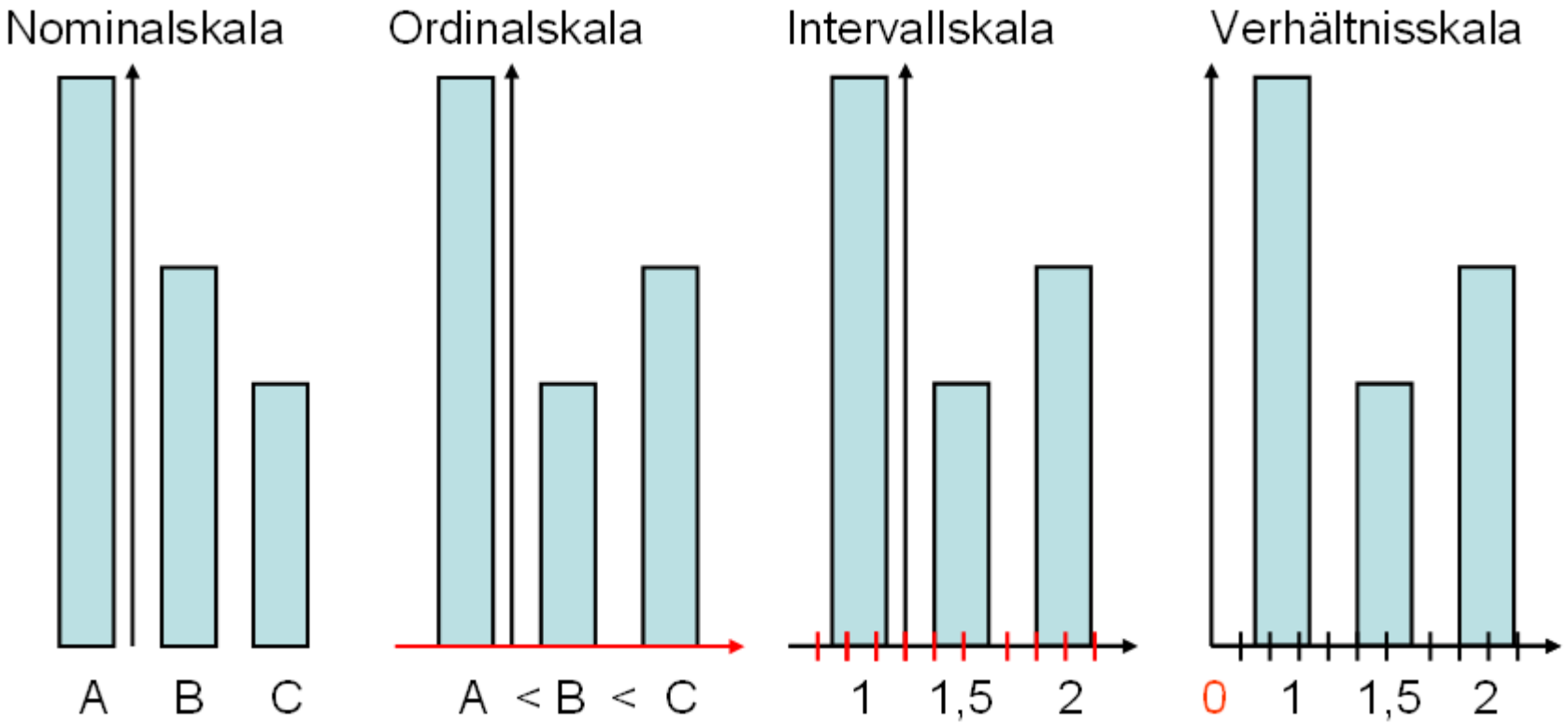
metric:

- Variable has a defined system of measurement, all calculations are possible. To distinguish are

1. interval: The variable has an arbitrary choosen neutral point (°C)
2. ratio: The variable has an absolute neutral point (°K)

- Sometimes also used: absolut scale
  - counts (number of inhabitans)

Levels of measurement



continuous vs. discrete

discrete variable:

- Variable which can take only certain values without intermediate values
- e.g. income, counts of ceramic objects, sex (?)
- 'counted'

continuous variables:

- Variable which can take all value and intermediate value
- e.g. height, temperature, proportion value
- 'measured'

Source: https://statsthewayilikeit.com

Cross tables (contingency tables)

For summary of data

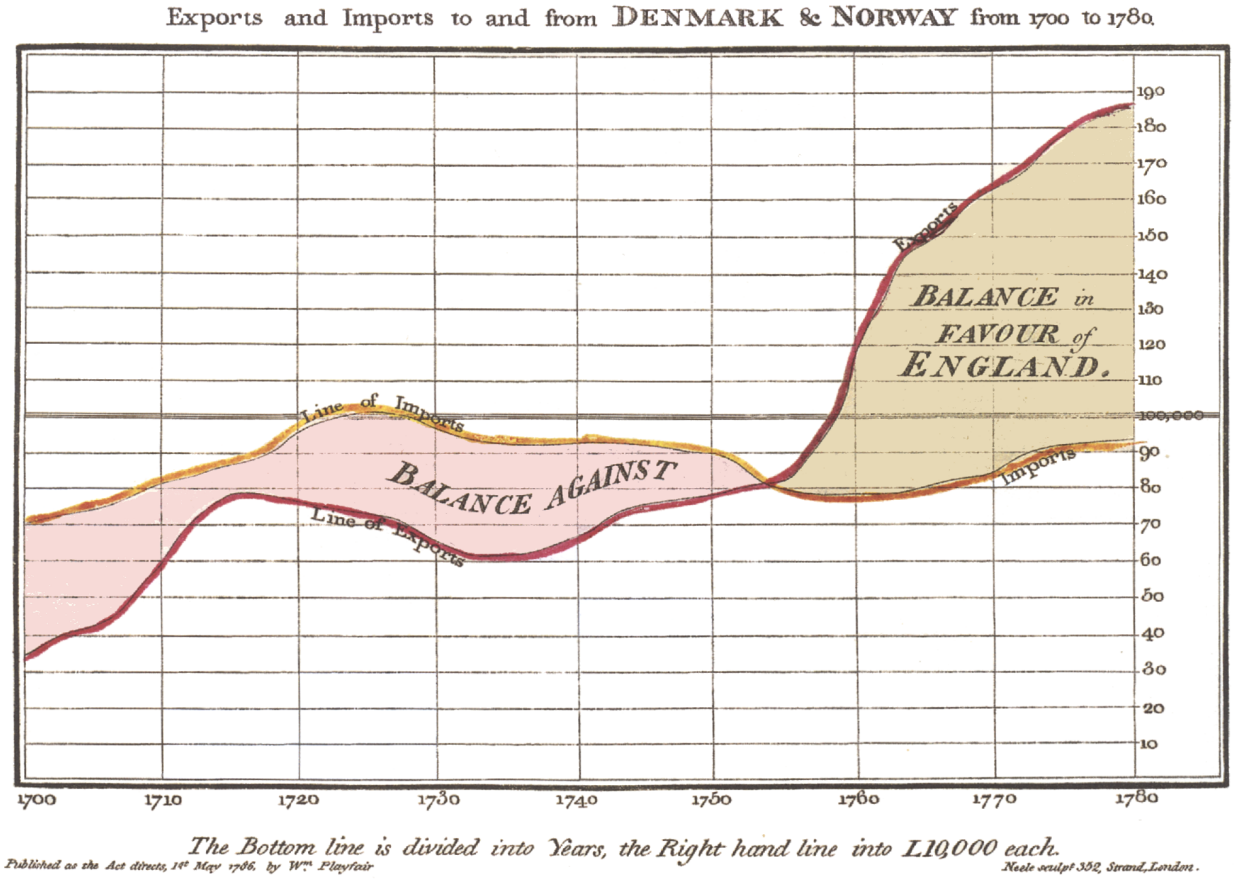
Cross tabulations summarise (mostly) categorical data by counting the co-occurrence of 2 (or more) categories per unit.

id	Schicht	Knubbenzahl
1	MSo	1
2	MSu	1
3	MSu	3
4	US	4
5	OSu	2
6	MSo	3
...	...	

Schicht	1 Knubbe	2 Knubben	3 Knubben	...
MSu	1	0	1	
MSo	1	0	0	
Osu	0	1	0	
...				

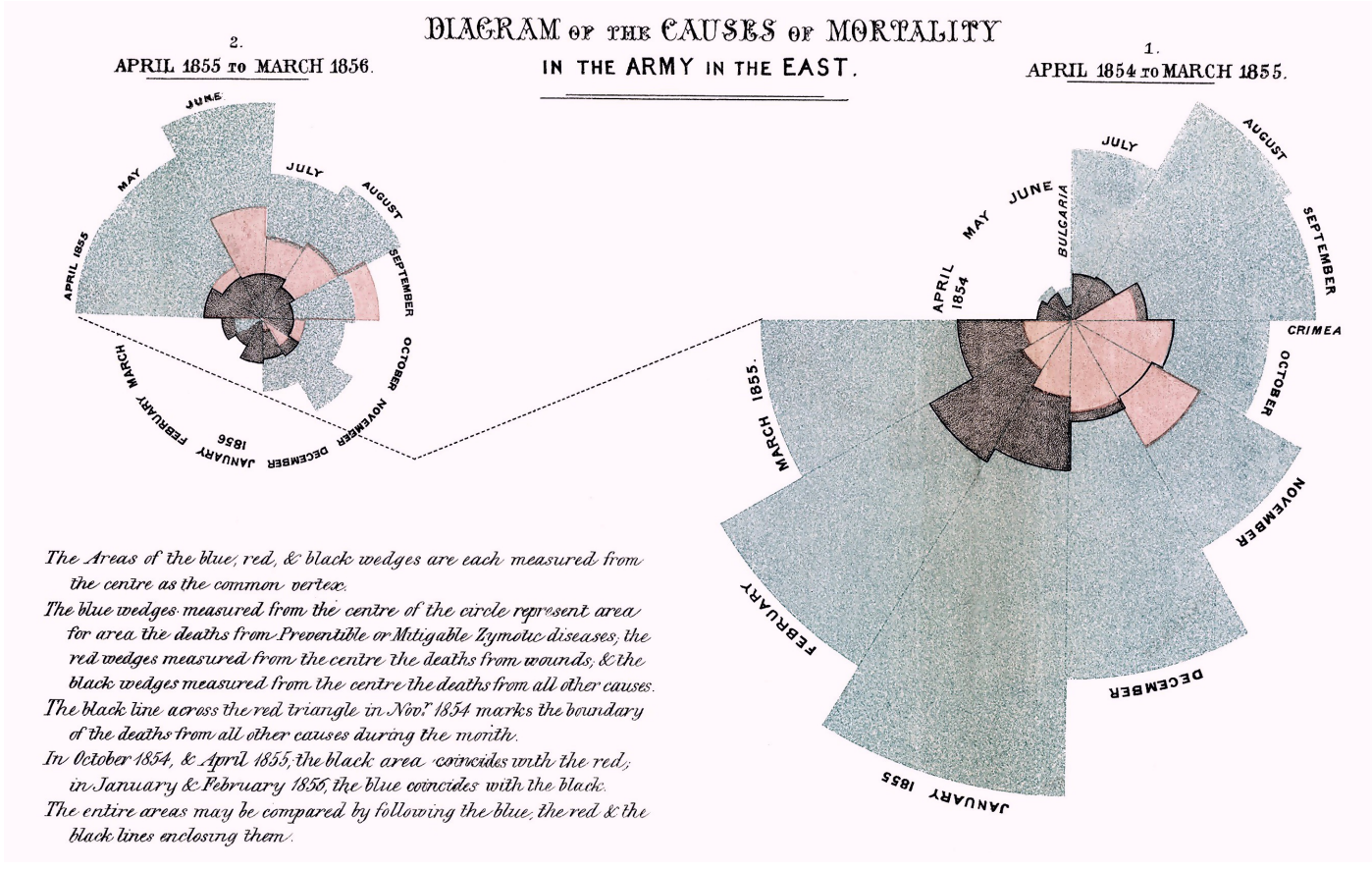
-> "Pivot-Table"

Successful visualisation of the past



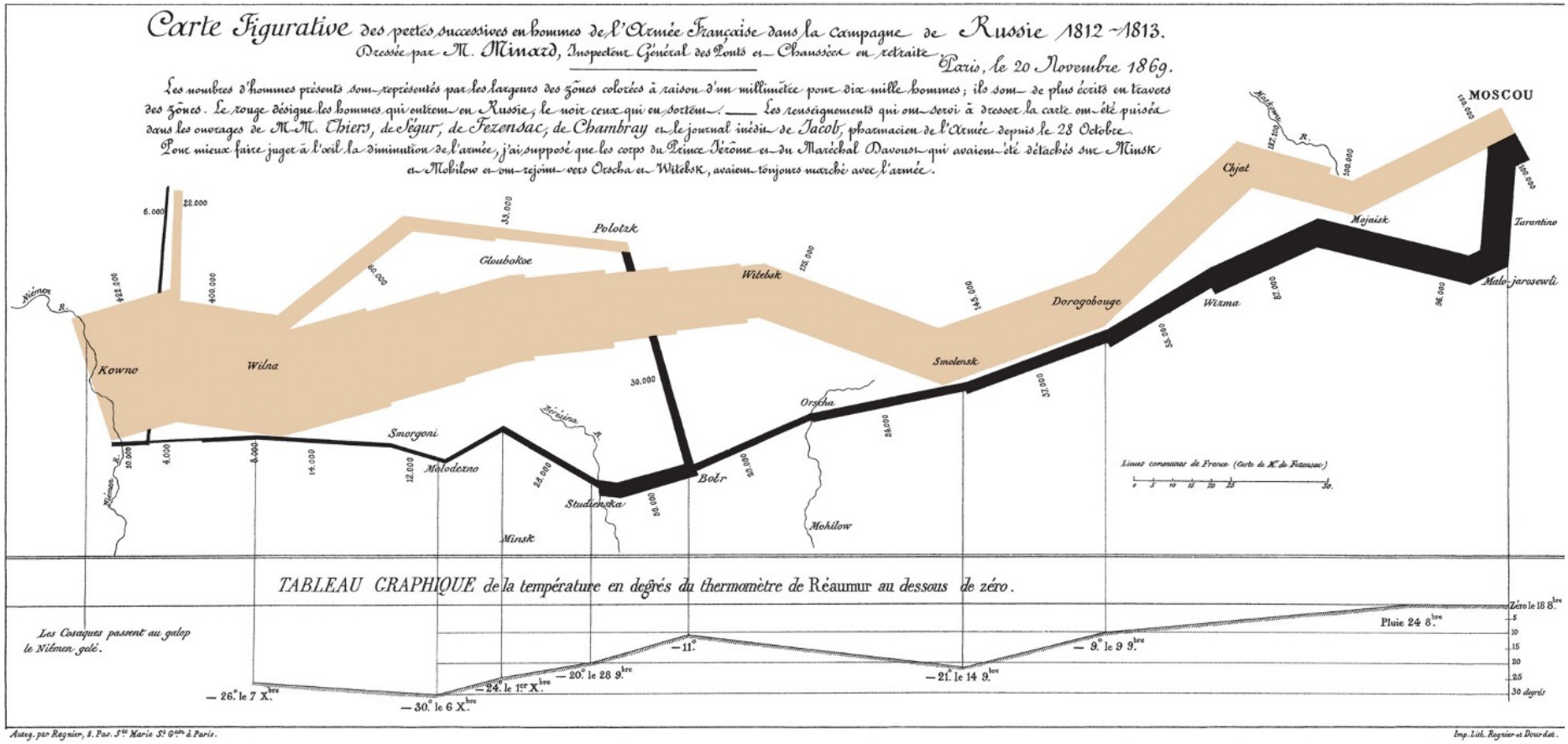
William Playfair, 1786. source: wikipedia

Successful visualisation of the past



Florence Nightingale, 1857. source: wikipedia

Successful visualisation of the past

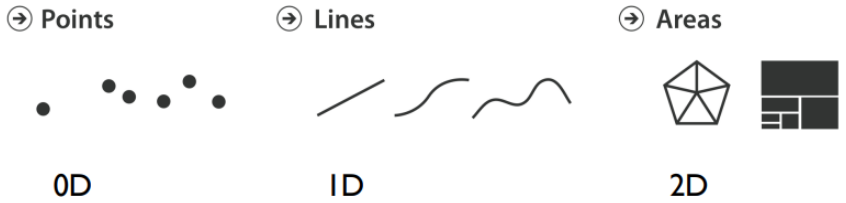


Charles Joseph Minard, 1869. source: wikipedia

Objects of Visualisation

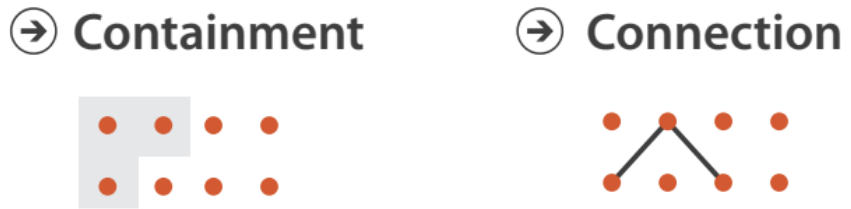
Items

Objects you like to display (entities in DB speach)



Links

The relationships of these objects, if there are any (Relationship also in DB speach)



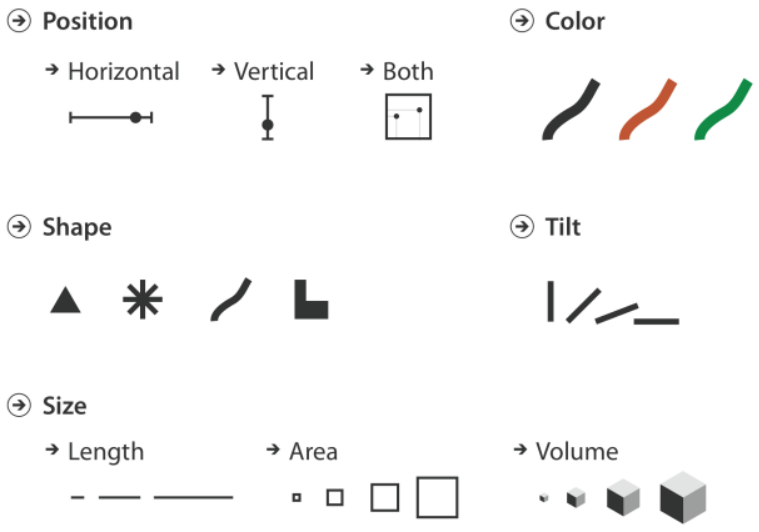
Machiraju 2020

Dimensions of Visualisation

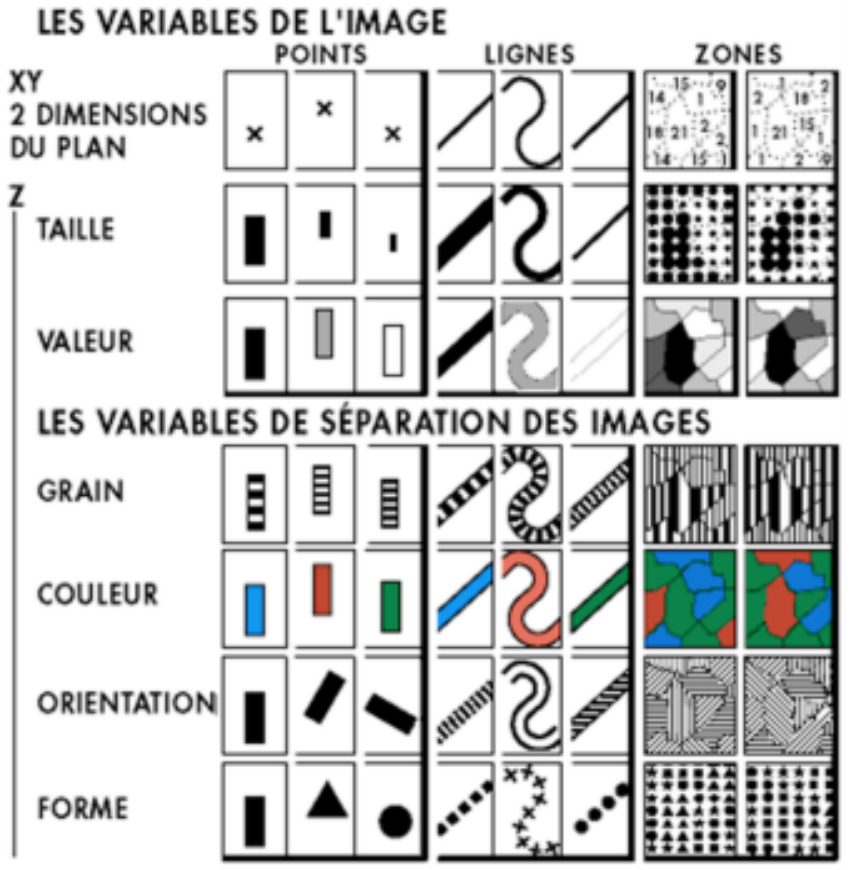
- Position (coordinates, slope, orientation, ...)
- Color (Hue, Saturation, Transparency)
- Texture
- Shape

- Size (length, area, [Volume])

- Proximity/Density



Machiraju 2020



Bertin 1967

Combining Dimensions of Visualisation

Mark: Line

Channel: Length/Position

1 quantitative attribute

1 categorical attribute

Mark: Point

Channel: Position

2 quantitative attr.

Adding Hue

+1 categorical attr.

Adding Size

+1 quantitative attr.

Machiraju 2020

Dimensions of Visualisation and levels of measurement

- Position (coordinates, slope, orientation, ...)
- Color (Hue, Saturation, Transparency)
- Texture
- Shape

- Size (length, area, [Volume])

- Proximity/Density

property	marks	ordinal/nominal mapping	quantitative mapping
shape	glyph	○ □ + △ S U	
size	rectangle, circle, glyph, text		
orientation	rectangle, line, text		
color	rectangle, circle, line, glyph, y-bar, x-bar, text, gantt bar		min max

Machiraju 2020

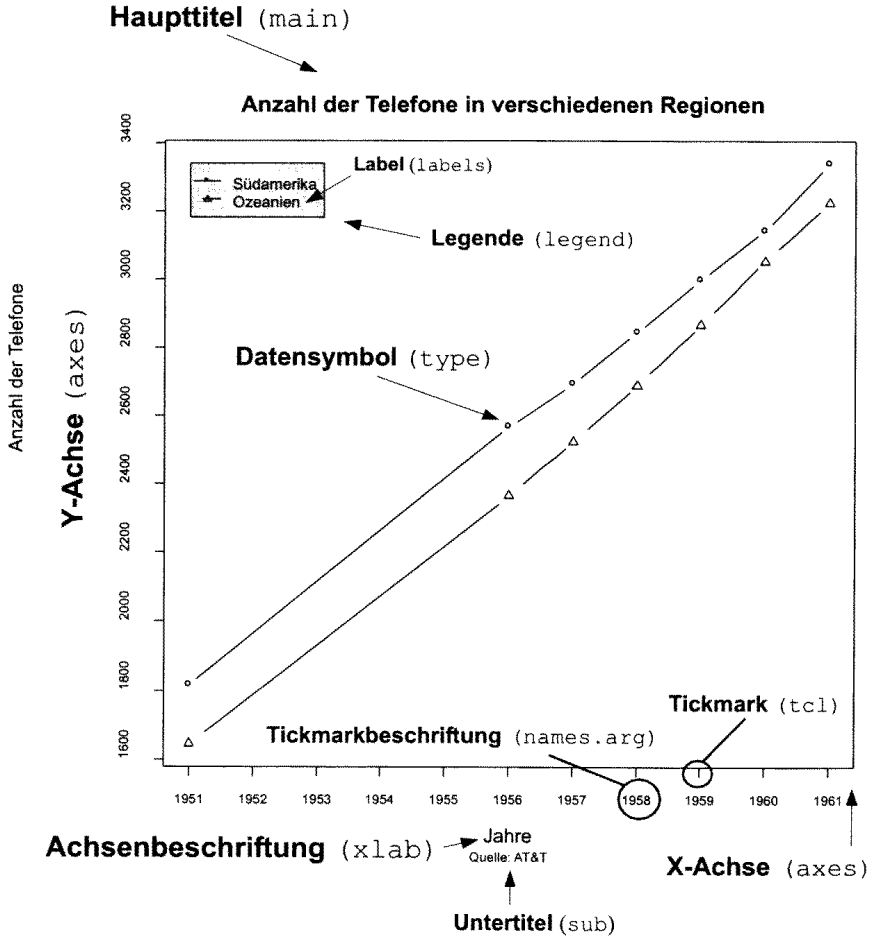
Basics about charts

Principles for good charts according to E. Tufte:

(The Visual Display of Quantitative Information. Cheshire/ Connecticut: Graphics Press, 1983)

- „Graphical exellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space.”
- Data-ink ratio = „proportion of a graphic’s ink devoted to the non-redundant display of data-information“ (kein chartjunk!)
- „Graphical excellence is often found in simplicity of design and complexity of data.“

- after Müller-Scheeßel



Pie chart [1]

The classical one – but comes with destinct flaws...

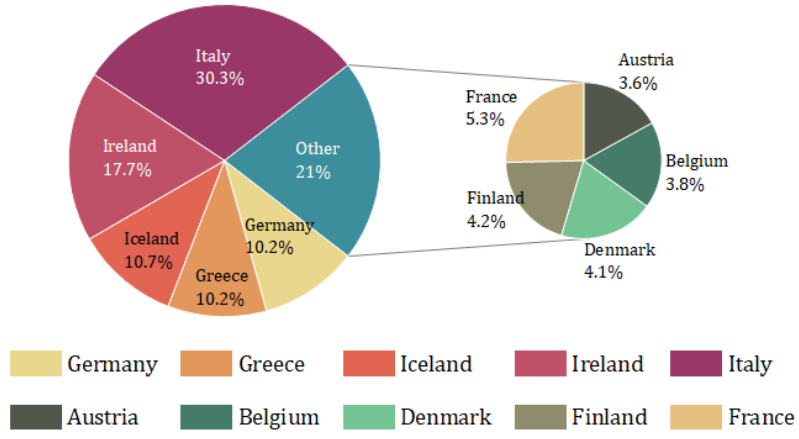
Used to display proportions, suitable for nominal data

$$a_i = \frac{n_i}{N} * 360^\circ$$

Disadvantages:

- Color selection can influence the perception (red is seen larger then gray)
- Small differences are not easy visible

totally No-Go: 3d-pies!!!



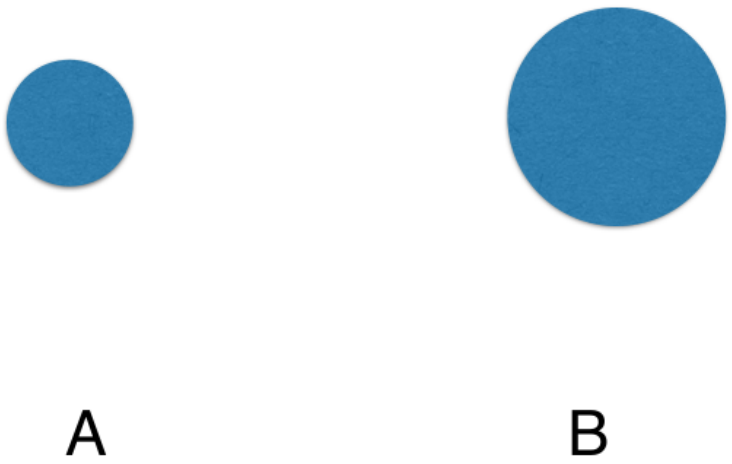
Pie chart [2]

I eat pie...

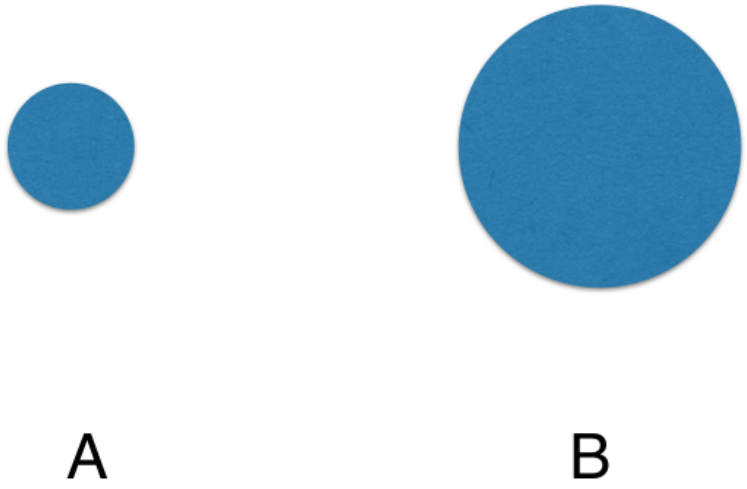
The pieces »viel zu wenig«, »etwas zu wenig« und »gerade richtig« have exactly the same size, the piece »viel zu viel« is a bit smaller.

source: http://www.lrz-muenchen.de/~wlm

Pie chart [3]

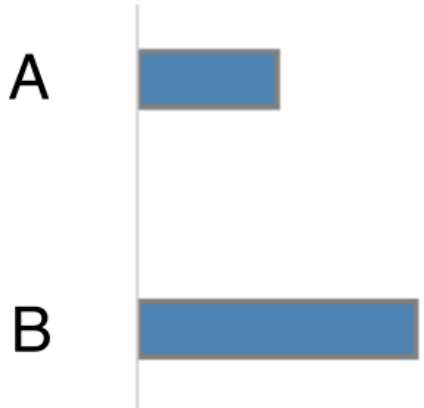


3x

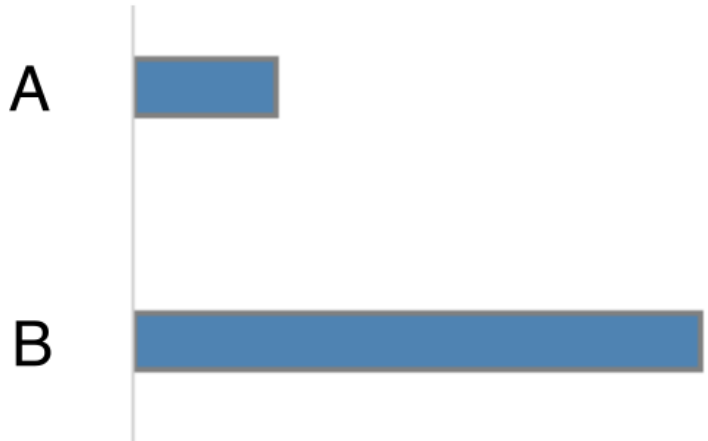


5x

Bar plot [1]



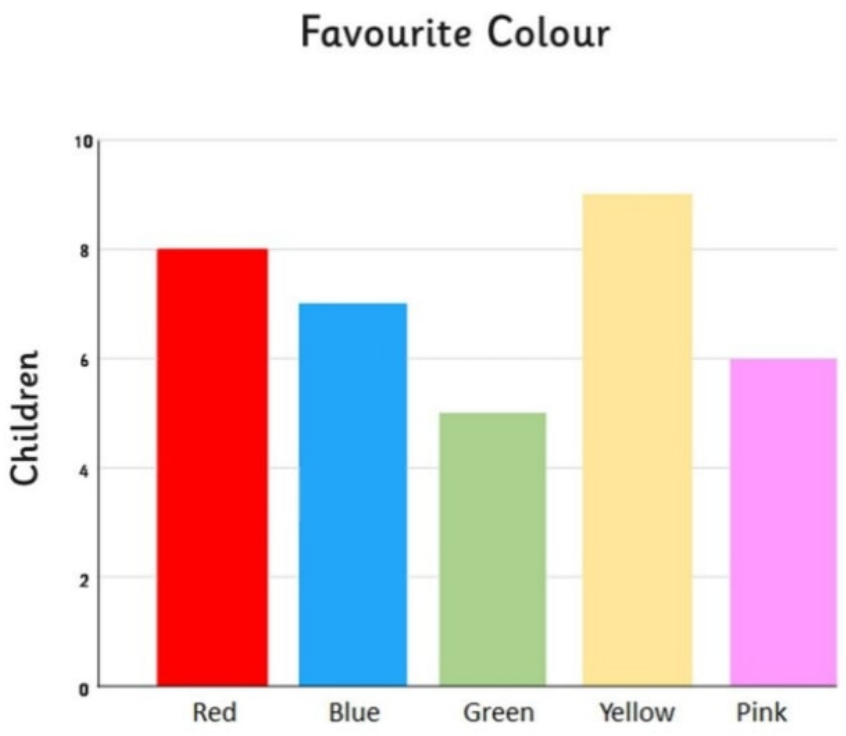
2x

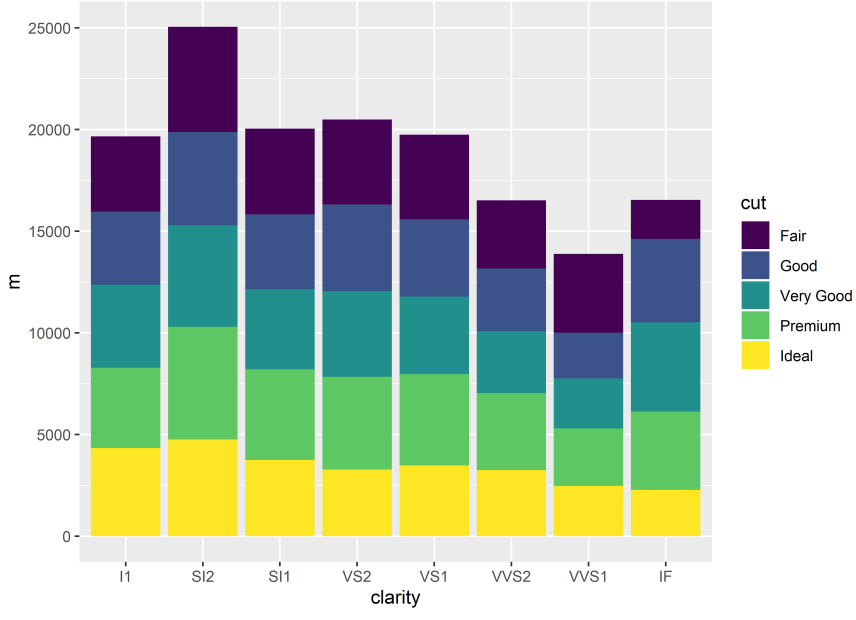


4x

Bar plot [2]

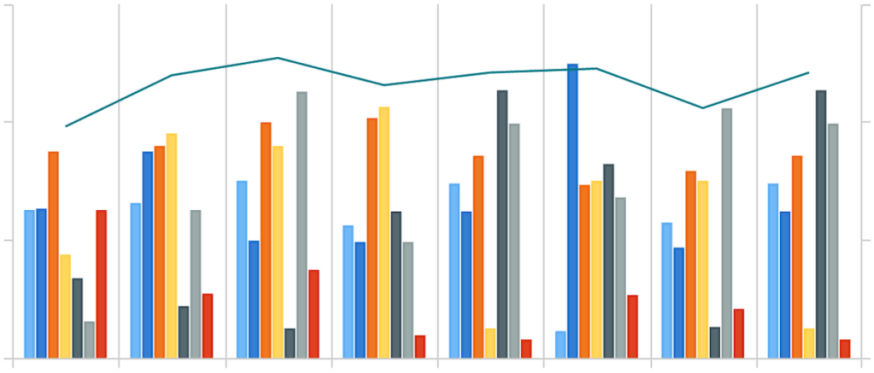
Generally the better alternative... Bar plots are suitable for display of proportions as well as for absolute data. They can be used for every level of measurement.

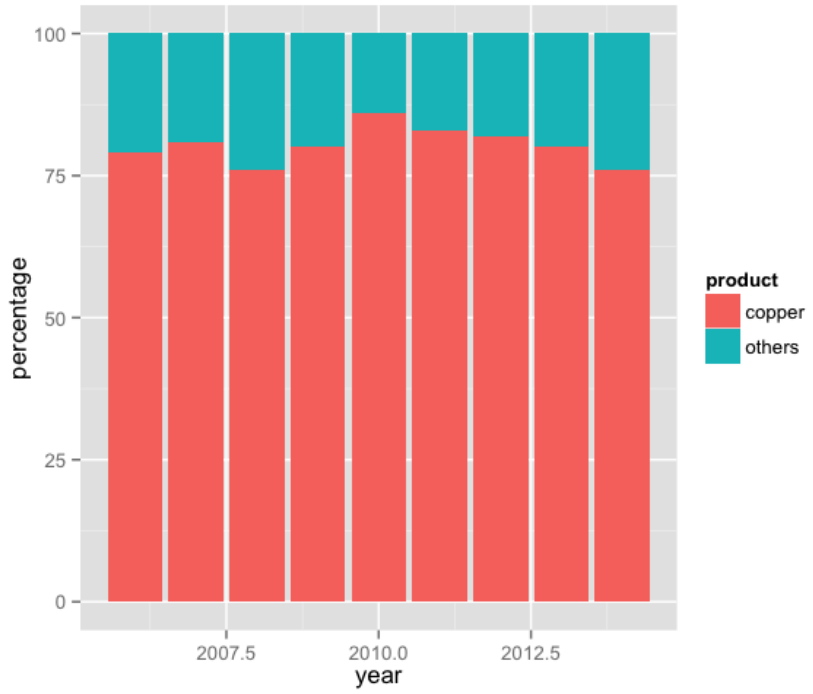




Bar plot [3]

Combination of different information and proportional visualisation is possible.



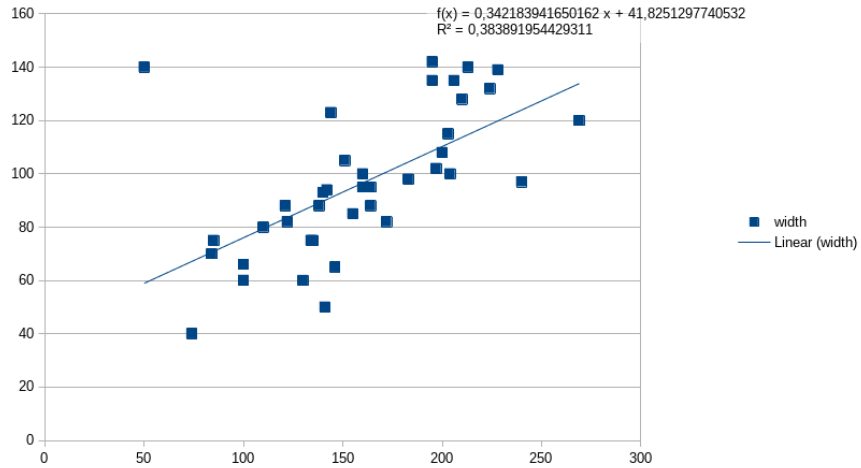


Scatterplot

Shows the relationship between two (metric) variables

You can see:

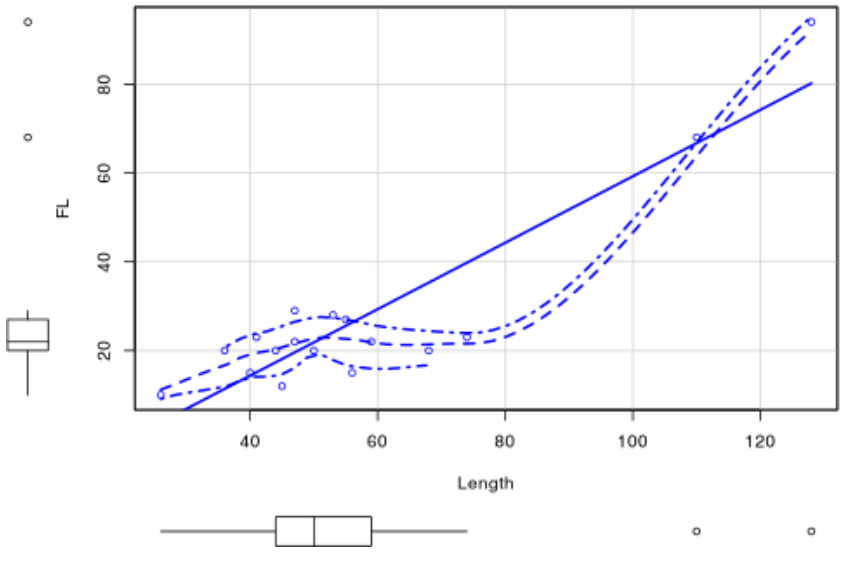
- values of items (points) on both variables
- relationship between variables
  - positive or negative relationship (or no at all)
- you can compute quantitative values describing the relationship
  - regression analysis



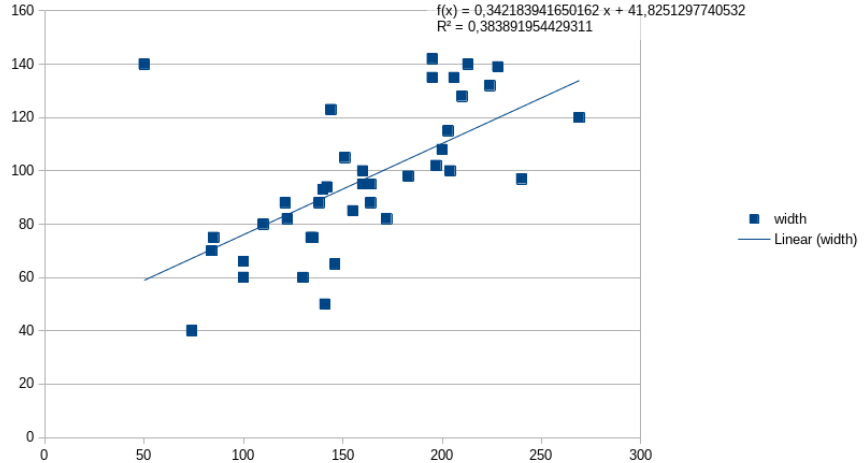
Scatterplot

Shows the relationship between two (metric) variables

R



Libre Office



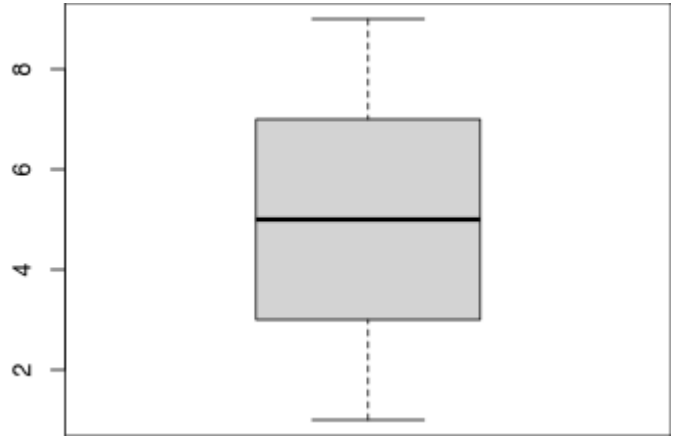
Box-plot (Box-and-Whiskers-Plot)

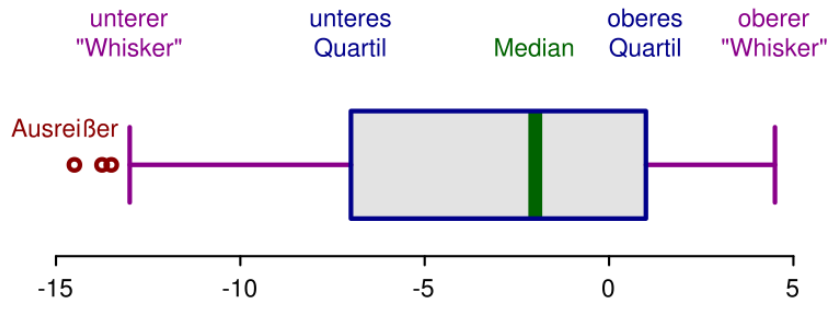
One of the best (my precious)!

Used to display the distribution of values in a data vector of metrical (interval, ratio) scale

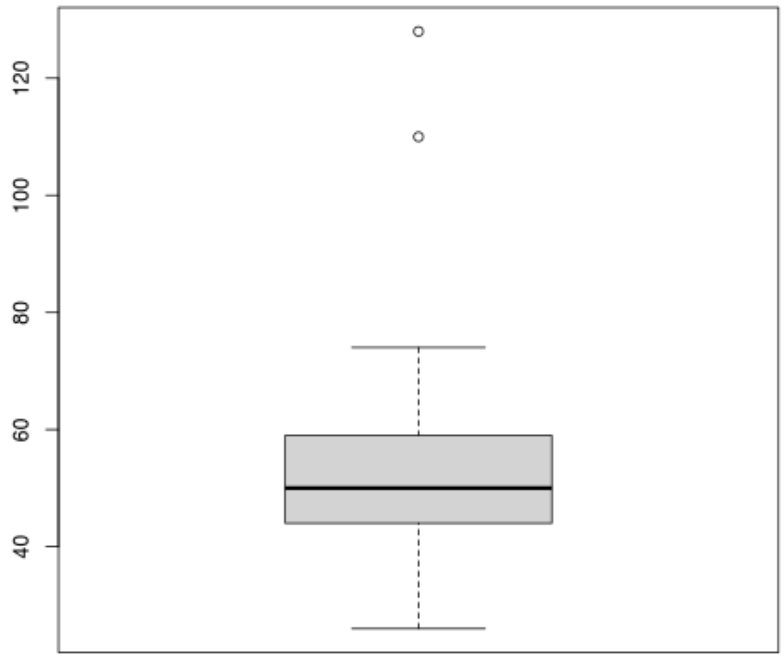
```
1  2  3  4  5  6  7  8  9
____|___|___|____
```

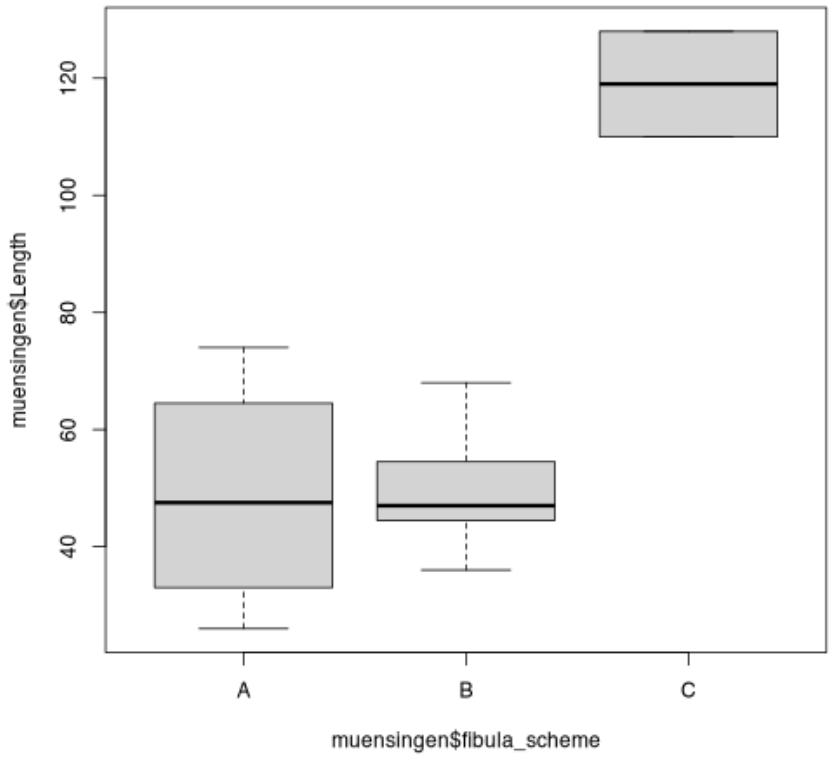
- thick line: median
- Box: the inner both quantiles
- Whisker: last value < than 1.5 times the distance of the inner quantile





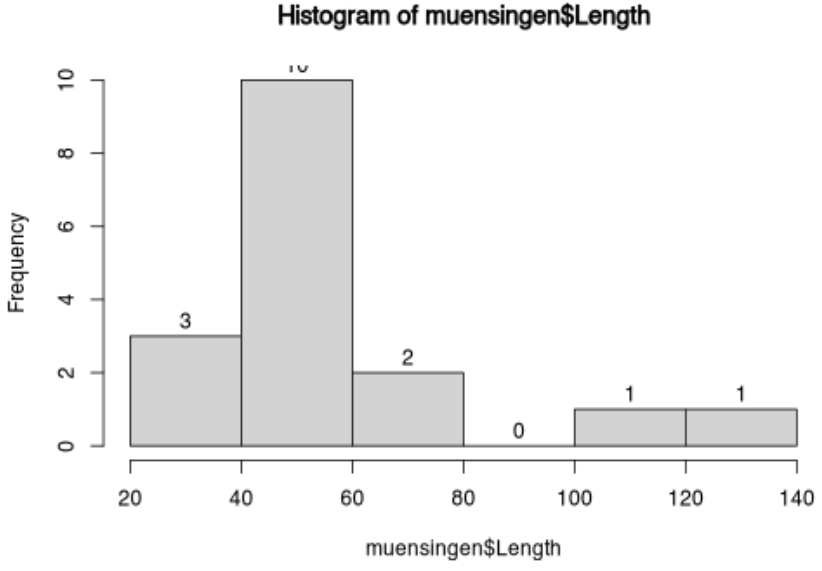
Box Plot [2]


120
100
80
60
40


muensingen$Length
120
100
80
60
40
A
B
C
muensingen$fibula_scheme

Histogramm

Used for classified display of distributions Data reduction vs. precision: Display of count values of classes of values



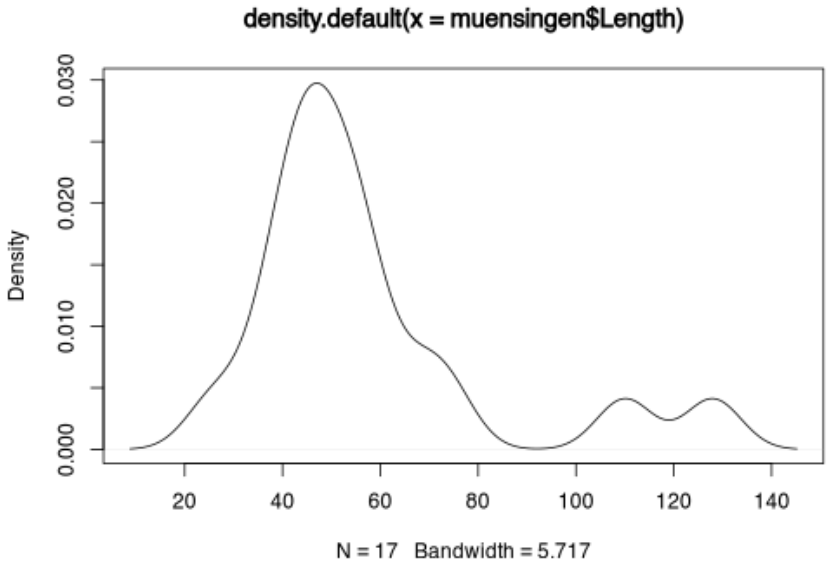
Disadvantages:

- Data reduction vs. precision → loss of information
- Actual display depends strongly on the choosen class width

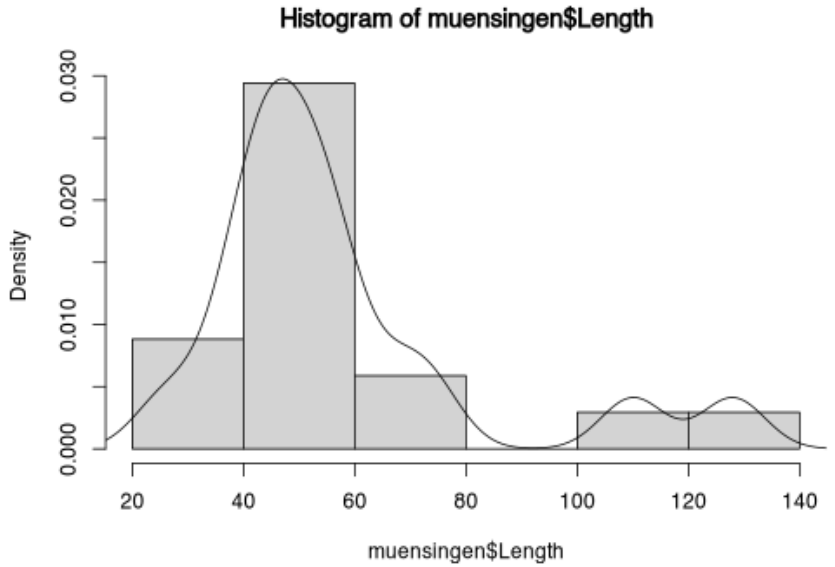
kernel smoothing (kernel density estimation)

Another attempt to overcome the disadvantages of a histogram

The distribution of the values is considered and a distribution curve is calculated. Continuous distributions are better displayed, without artificial breaks. Scales like histograms.



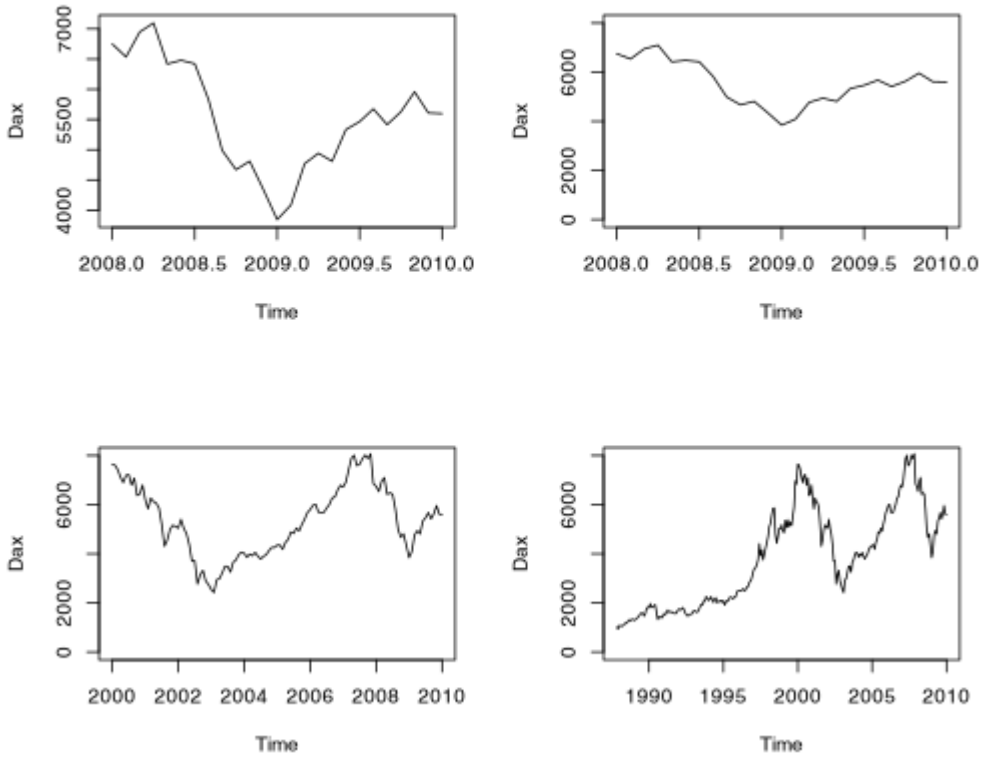
Histogram and kernel-density-plot together



Style of charts

Stay honest!

Choice of display has a strong influence on the statement.



Style of charts

Stay honest!

Choice of display has a strong influence on the statement.

Clear layout!

Minimise Ratio of ink per shown information!

Use the suitable chart for the data!

Consider nominal-ordinal-interval-ratio scale

Suggestions for charts

What to display	suitable	not suitable
Parts of a whole: few	Pie chart, stacked bar plot	
Parts of a whole: few	Stacked bar plot	
Multiple answers (ties)	Horizontal bar plot	Pie chart, stacked bar plot
Comparison of different values of different variables	Grouped bar plot	
Comparison of parts of a whole	Stacked bar plot	
Comparison of developments	Line chart	
Frequency distribution	Histogram, kernel density plot	
Correlation of two variables	scatterplot	

Any questions?

You might find the course material (including the presentations) at

https://berncodalab.github.io/caa

You can contact me at

martin.hinz@iaw.unibe.ch