

Computer Aided Archaeology

05 - Visualisation I

Martin Hinz

Institut für Archäologische Wissenschaften, Universität Bern

18/10/23

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



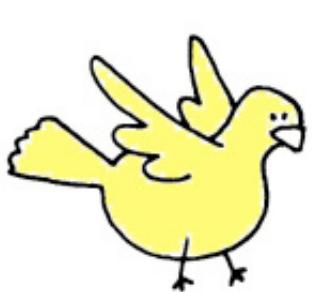
Data, variables, values

- 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.

The screenshot shows a CSV file named "kursdata.csv" open in a spreadsheet program. The data consists of 15 rows of information about individuals, with columns labeled A through H. Row 1 contains column headers: A (empty), B (age), C (height), D (sex), E (empty), F (empty), G (empty), and H (empty). Rows 2 through 15 contain data for 14 individuals. Column B contains names like Hannah, Leon, Lukas, Leonie, Luka, Lea, Lena, Mia, Tim, Lynn, Anna, Emily, Felix, and Martin. Column C contains numerical values for height. Column D contains 'f' for female and 'm' for male. Handwritten annotations include a red arrow pointing to column D labeled "Variable", a red arrow pointing to row 16 labeled "Item", and another red arrow pointing to the value "431,81" in cell D15 labeled "Value".

A	B	C	D	E	F	G	H
1							
2	Hannah	221,68	f				
3	Leon	251,67	m				
4	Lukas	201,87	m				
5	Leonie	241,65	f				
6	Luka	201,9	m				
7	Lea	221,76	f				
8	Lena	241,67	f				
9	Mia	211,56	f				
10	Tim	191,81	m				
11	Lynn	241,65	m				
12	Anna	251,67	f				
13	Emily	211,71	f				
14	Felix	181,54	m				
15	Martin	431,81	m				
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							

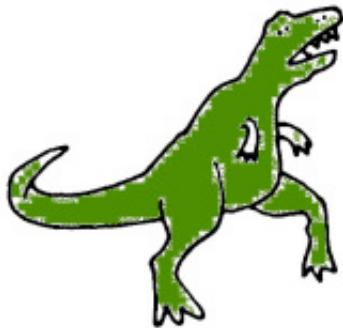
Levels of measurement



I am a bird.
I am yellow.
I am awesome.



I am a seahorse.
I am orange.
I am super awesome.

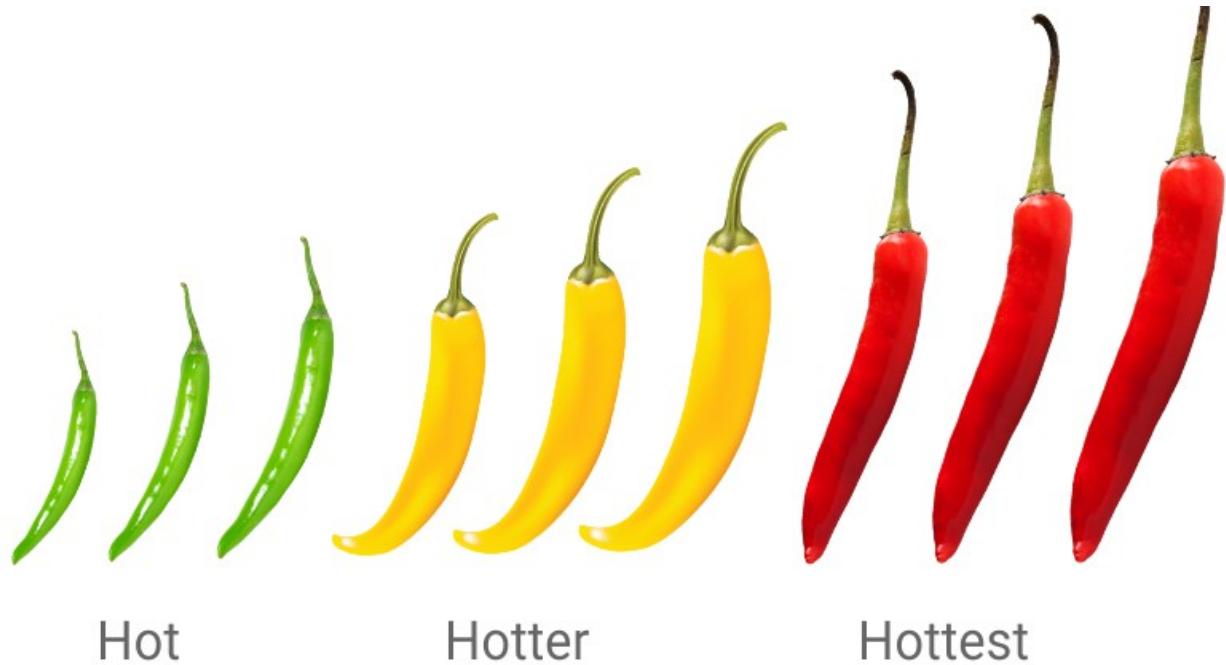


I am a T-rex.
I am green.
I am extinct.

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)

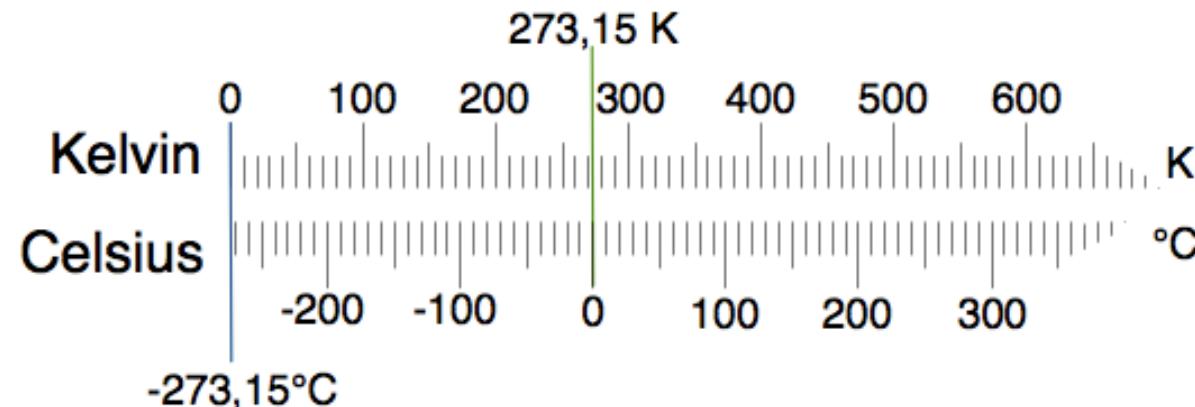
Levels of measurement



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

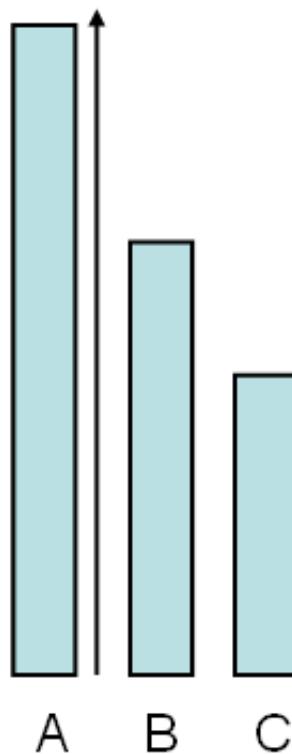


metric:

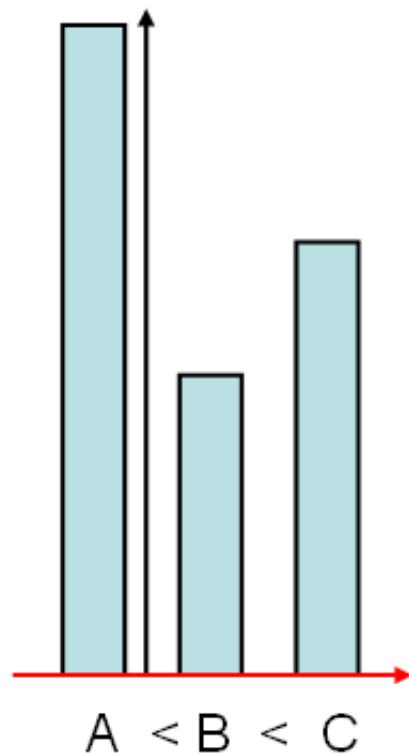
- Variable has a defined system of measurement, all calculations are possible. To distinguish are
 - 1. interval: The variable has an arbitrary chosen neutral point ($^{\circ}\text{C}$)
 - 2. ratio: The variable has an absolute neutral point ($^{\circ}\text{K}$)
- Sometimes also used: absolut scale
 - counts (number of inhabitants)

Levels of measurement

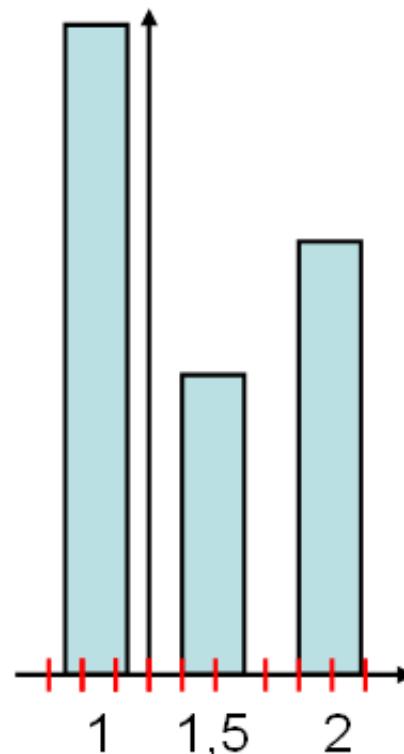
Nominalskala



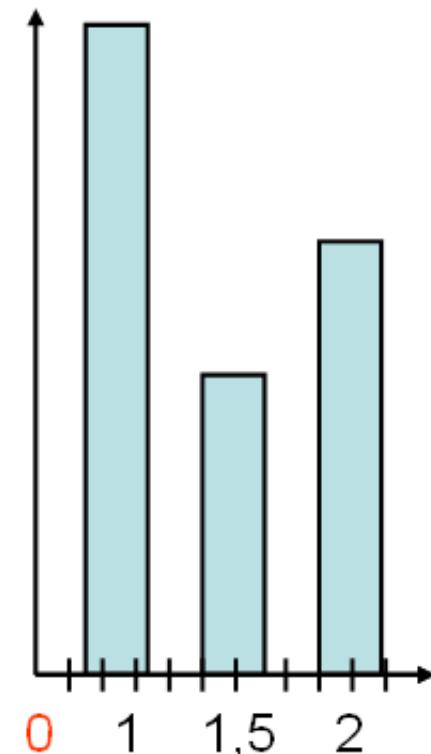
Ordinalskala



Intervallskala



Verhältnisskala



continuous vs. discrete

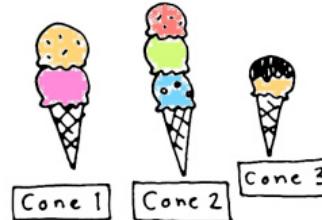
discrete variable:

- 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'

QUANTITATIVE DATA:



Discrete data:

- There are 3 cones
- Cone 1 has 2 scoops

Continuous data:

- Cone 3 weighs 79.4 grams
- cone 2 ice cream is at 8.3°F

Source: <https://statstthewayilikeit.com>

Cross tables (contingency tables)

For summary of data

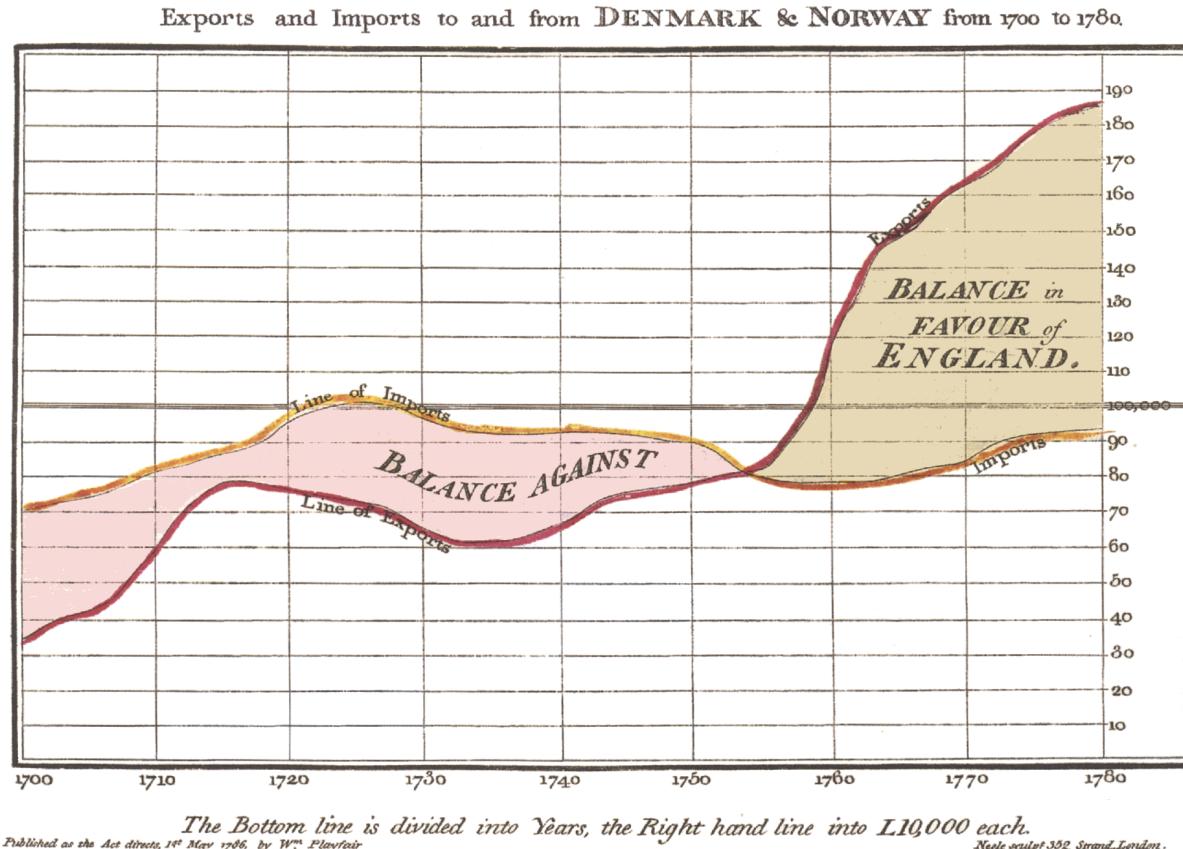
Cross tabulations summarise (mostly) categorical data by counting the co-occurrence of 2 (or more) categories per unit.

<u>id</u>	<u>Schicht</u>	<u>Knubbenzahl</u>
1	<u>MSo</u>	1
2	<u>MSu</u>	1
3	<u>MSu</u>	3
4	US	4
5	<u>OSu</u>	2
6	<u>MSo</u>	3
...	...	

<u>Schicht</u>	<u>1 Knubbe</u>	<u>2 Knubben</u>	<u>3 Knubben</u>	...
<u>MSu</u>	1	0	1	
<u>MSo</u>	1	0	0	
<u>Osu</u>	0	1	0	
...				

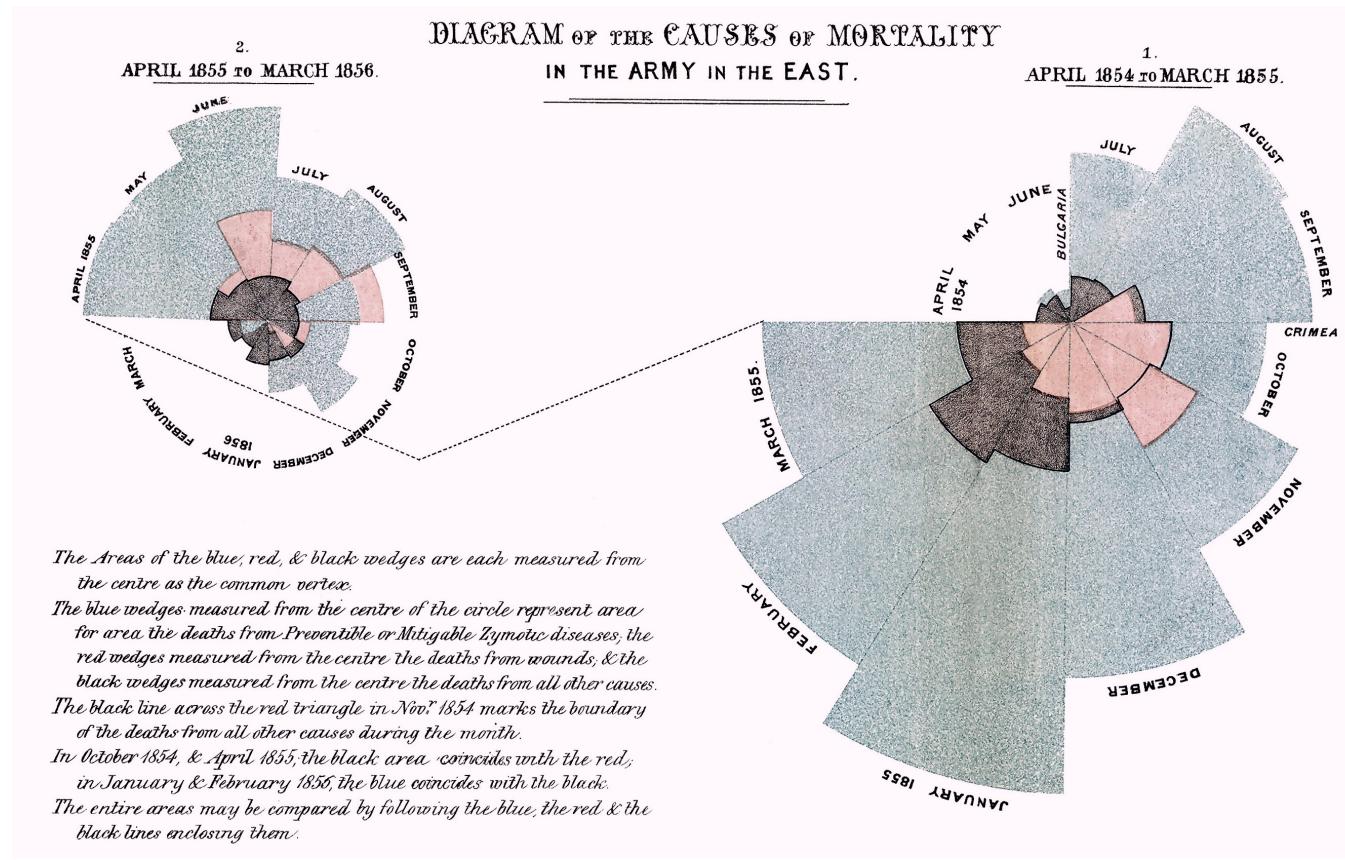
-> "Pivot-Table"

Successful visualisation of the past



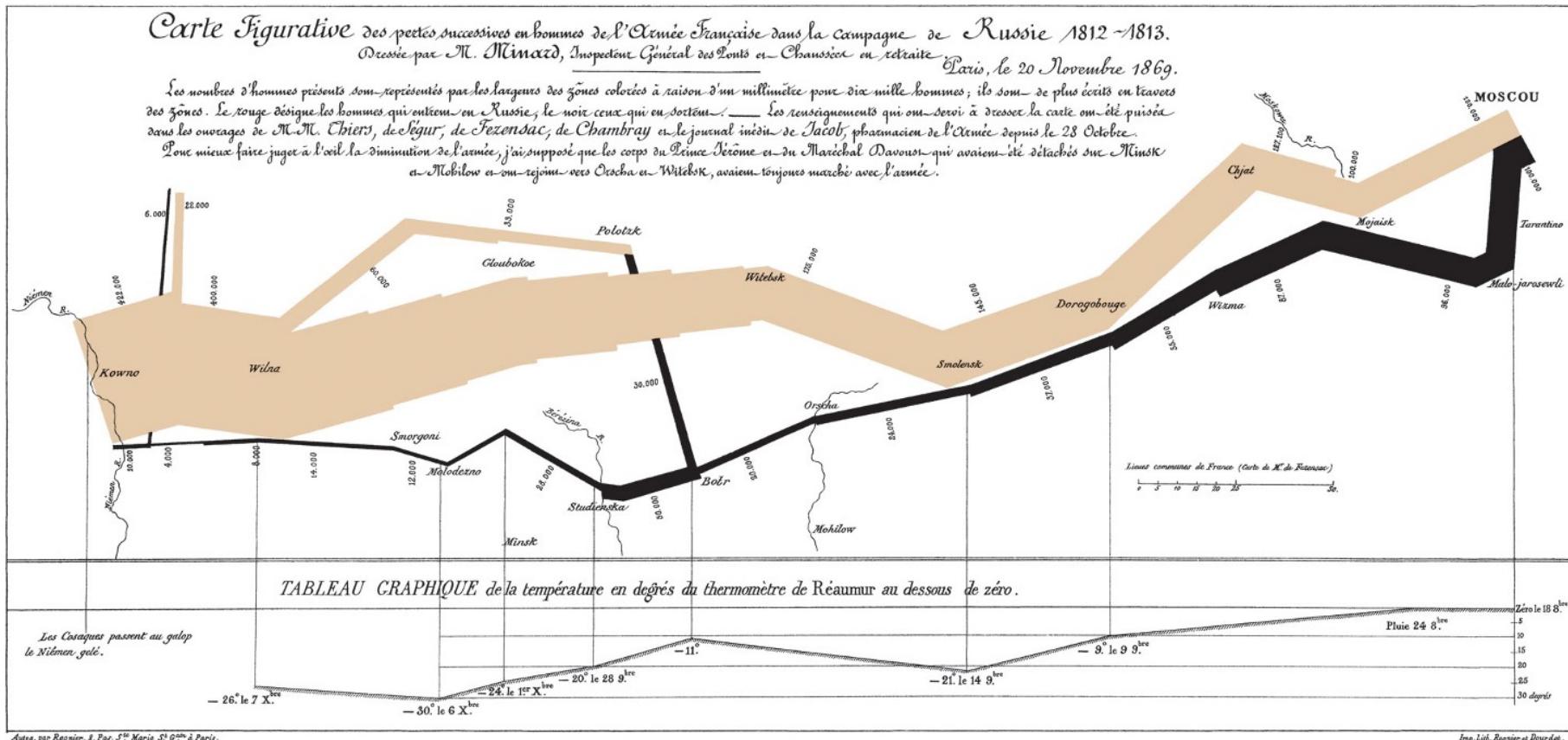
William Playfair, 1786. source: wikipedia

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 speak)

④ Points



0D

④ Lines



ID

④ Areas



2D

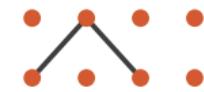
Links

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

➔ Containment



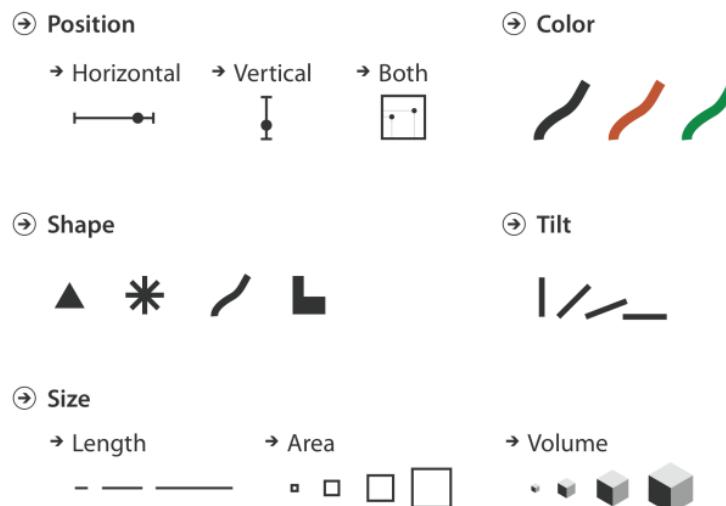
➔ Connection



Machiraju 2020

Dimensions of Visualisation

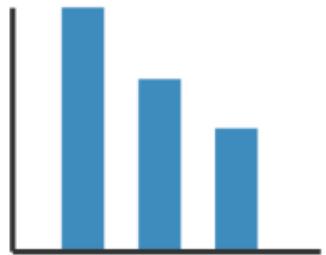
- Position (coordinates, slope, orientation, ...)
 - Color (Hue, Saturation, Transparency)
 - Texture
 - Shape
 - Size (length, area, [Volume])
 - Proximity/Density



LES VARIABLES DE L'IMAGE				
	POINTS	LIGNES	ZONES	
XY 2 DIMENSIONS DU PLAN				
Z TAILLE				
VALEUR				
LES VARIABLES DE SÉPARATION DES IMAGES				
GRAIN				
COULEUR				
ORIENTATION				
FORME				

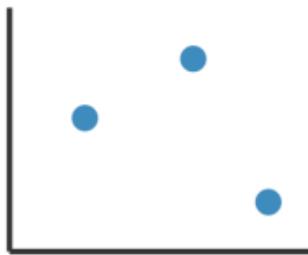
Bertin 1967

Combining Dimensions of Visualisation



Mark: Line

Channel: Length/Position
1 quantitative attribute
1 categorical attribute



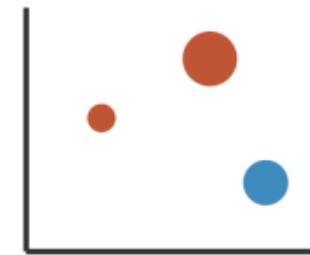
Mark: Point

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

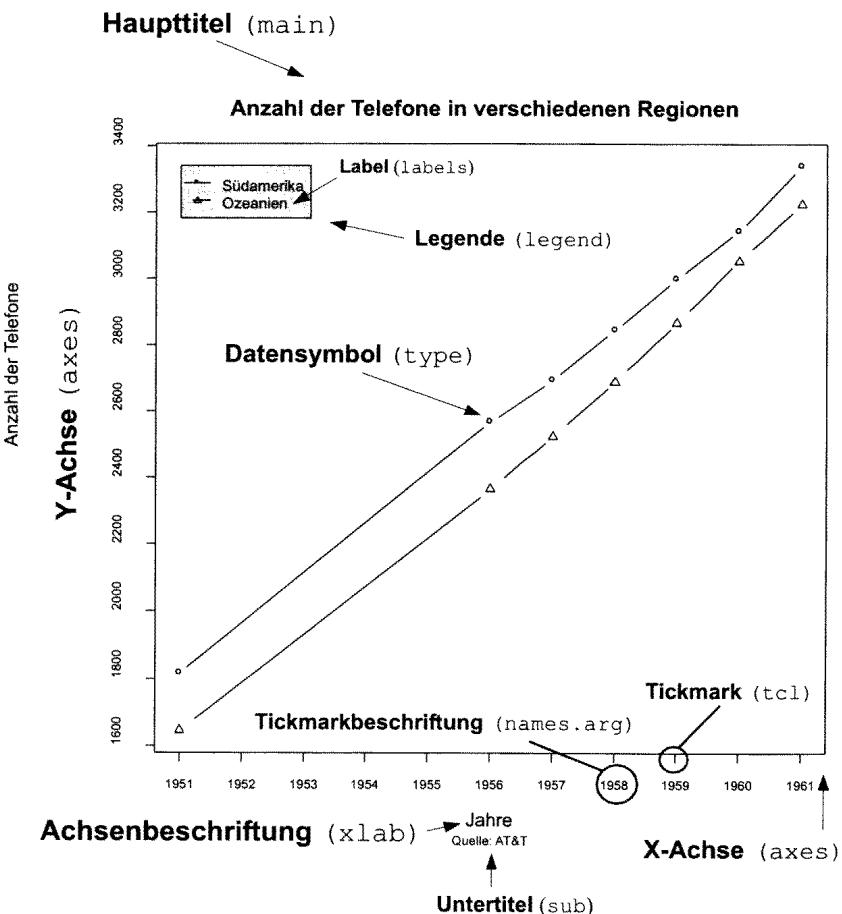
Basics about charts

Principles for good charts
according to E. Tufte:

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

- „Graphical excellence 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 distinct flaws...

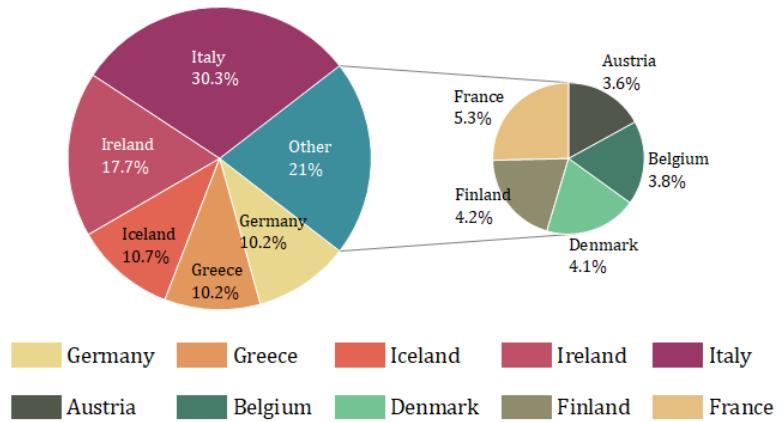
Used to display proportions, suitable for nominal data

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

Disadvantages:

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

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



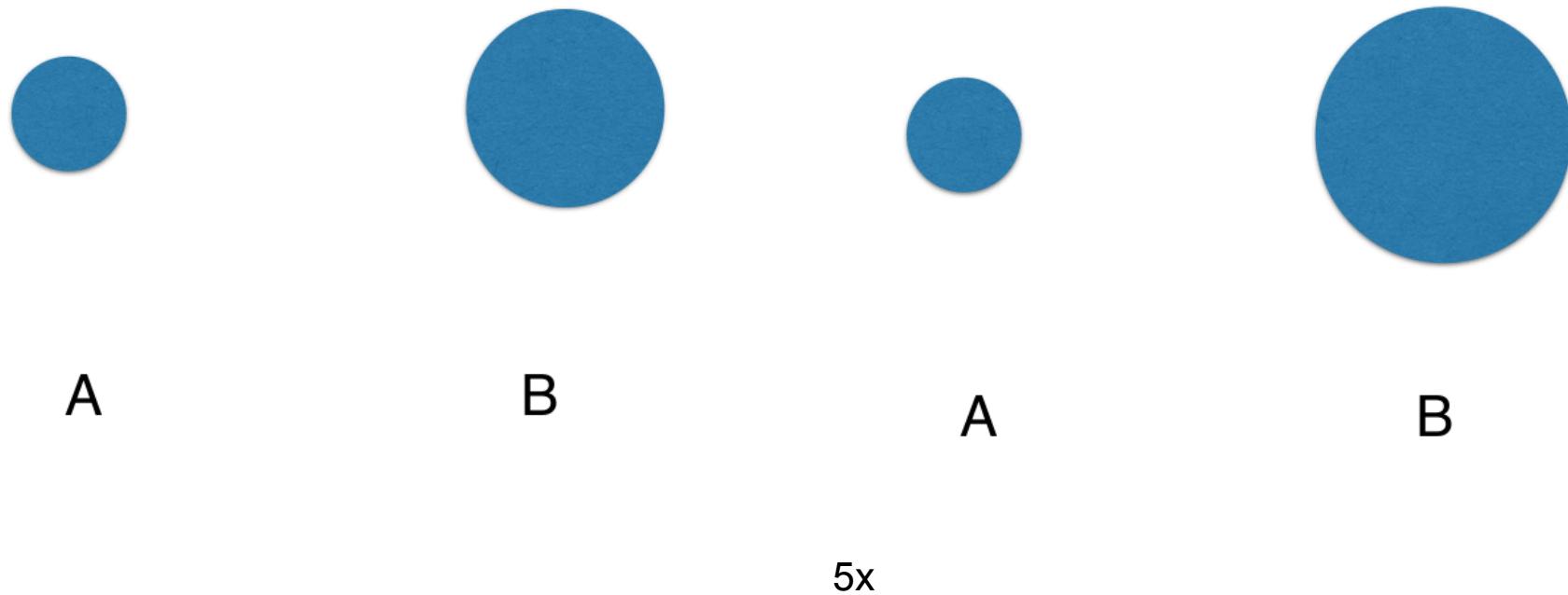
Pie chart [2]

I eat pie...

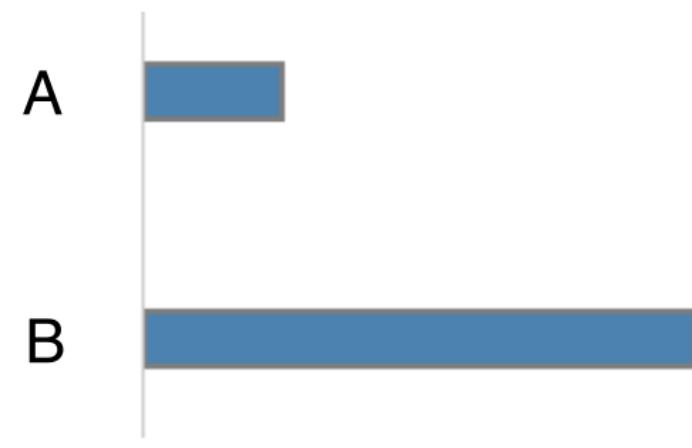
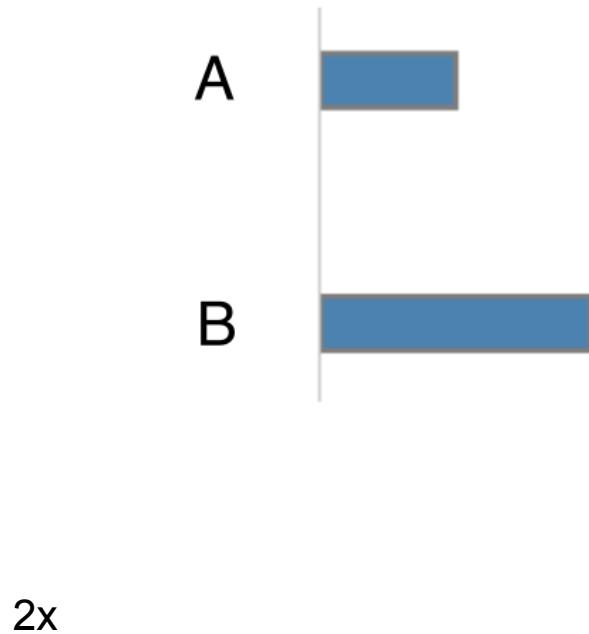


The pieces »viel zu wenig«, »etwas zu wenig« und »gerade richtig« have exactly the same size, the piece »viel zu viel« is a bit smaller. source: <http://www.lrz-muenchen.de/~wlm>

Pie chart [3]

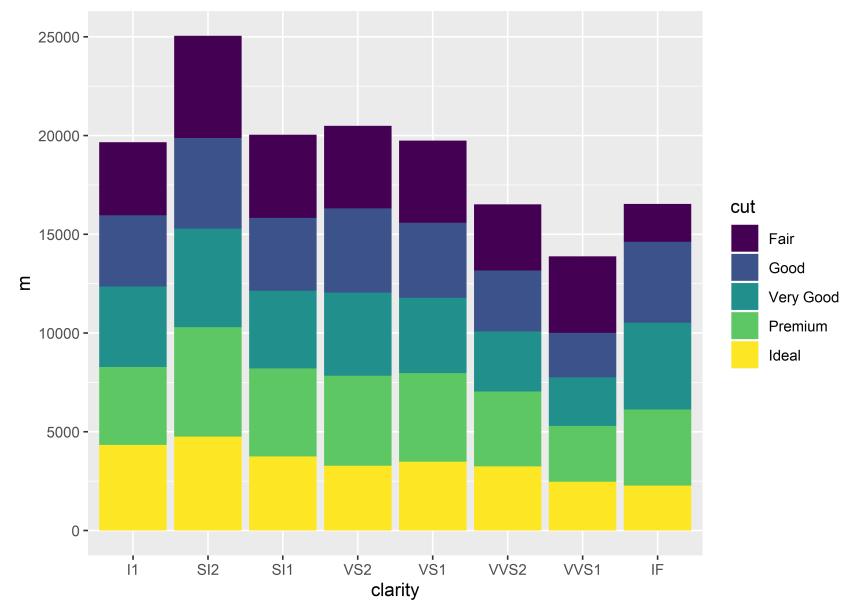
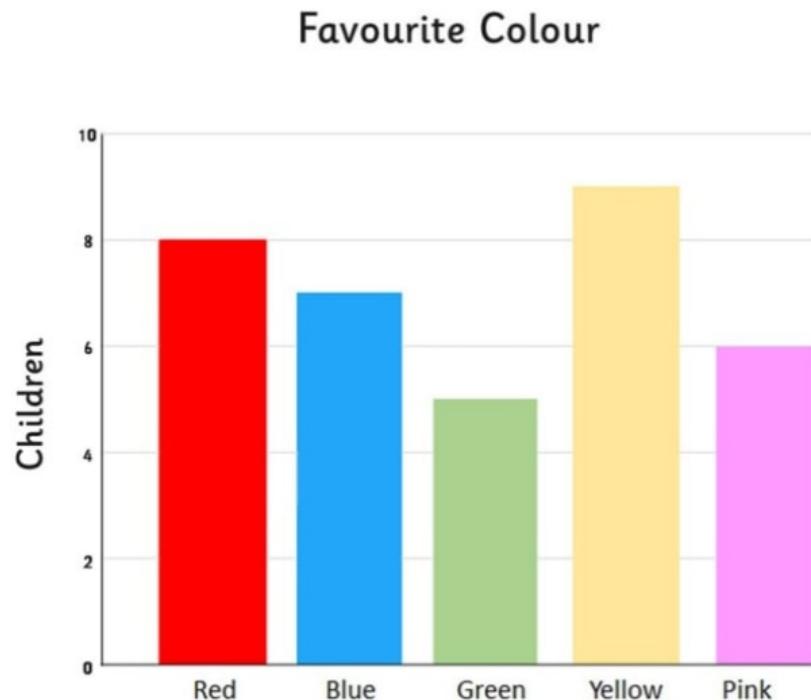


Bar plot [1]



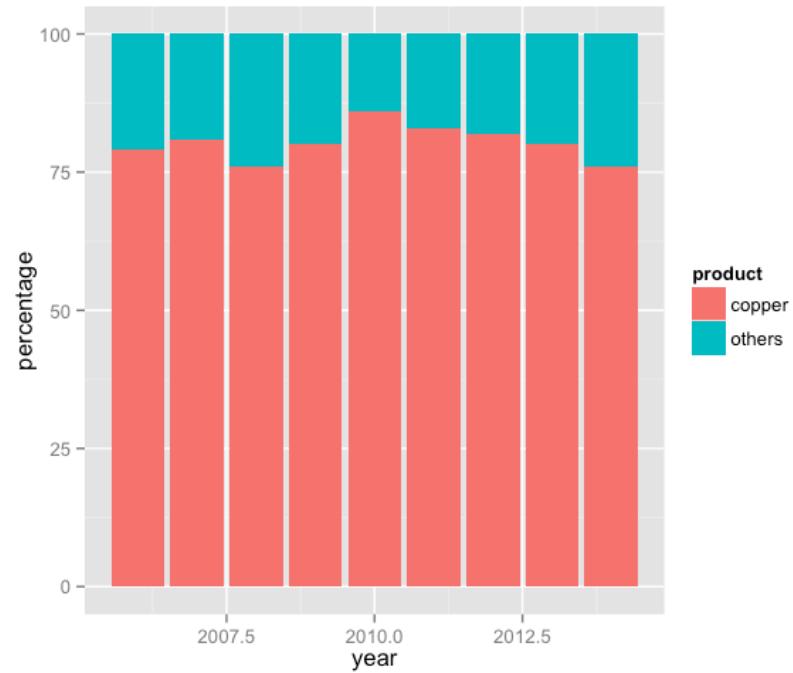
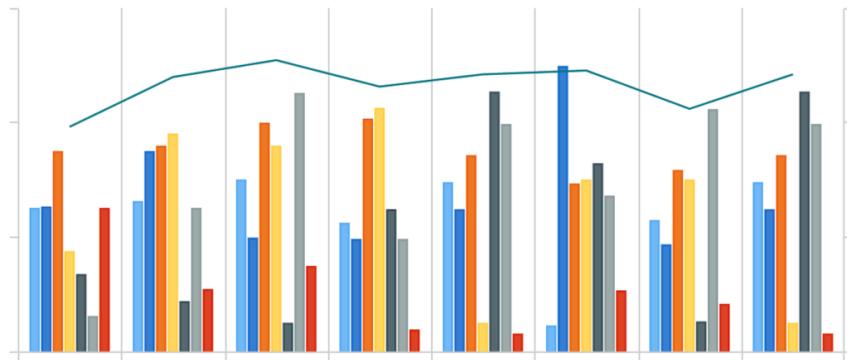
Bar plot [2]

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.

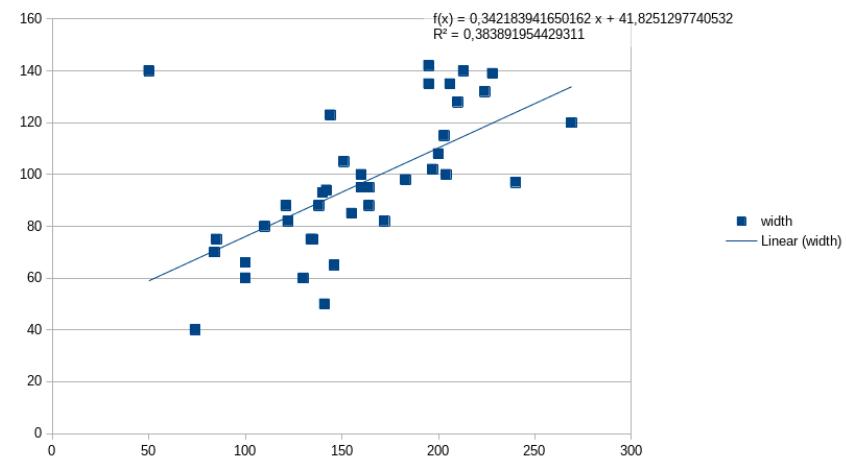


Scatterplot

Shows the relationship between two (metric) variables

You can see:

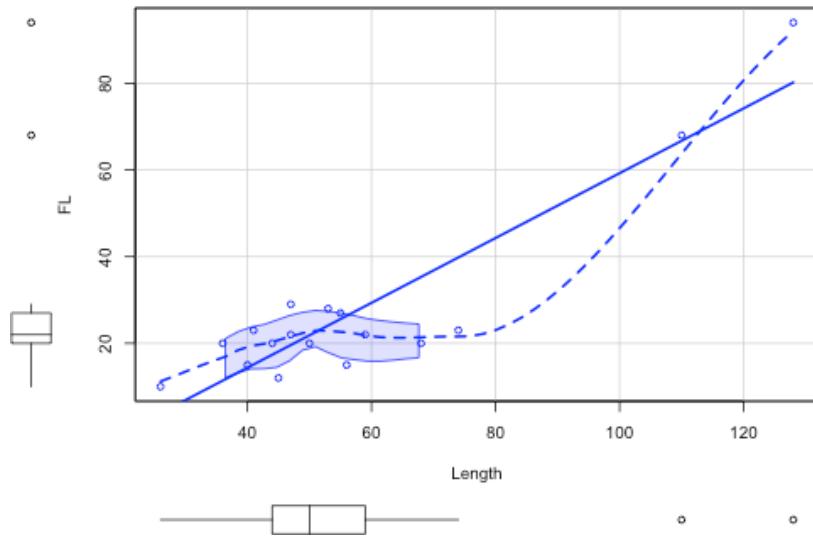
- values of items (points) on both variables
- relationship between variables
 - positive or negative relationship (or no at all)
- you can compute quantitative values describing the relationship
 - regression analysis



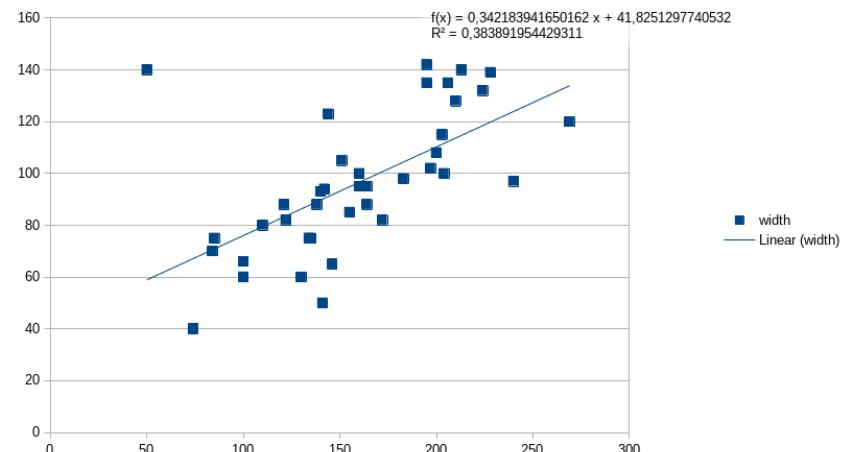
Scatterplot

Shows the relationship between two (metric) variables

R



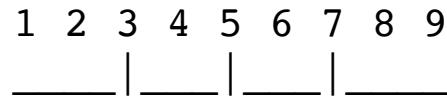
Libre Office



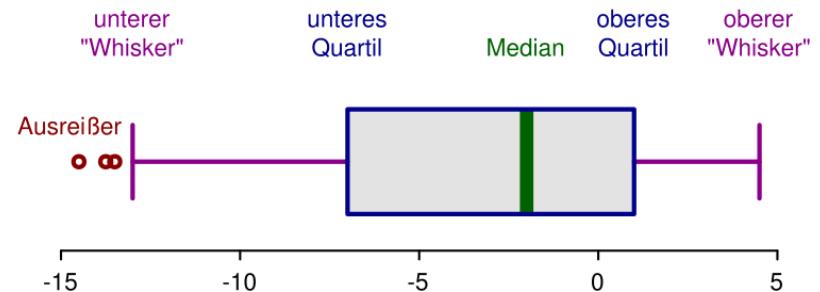
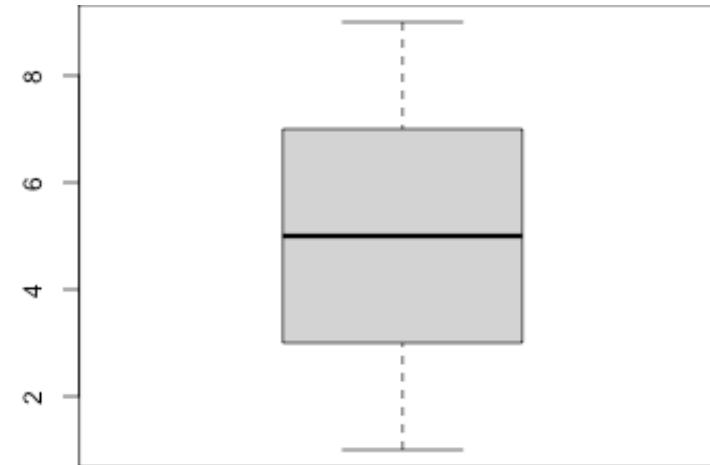
Box-plot (Box-and-Whiskers-Plot)

One of the best (my precious)!

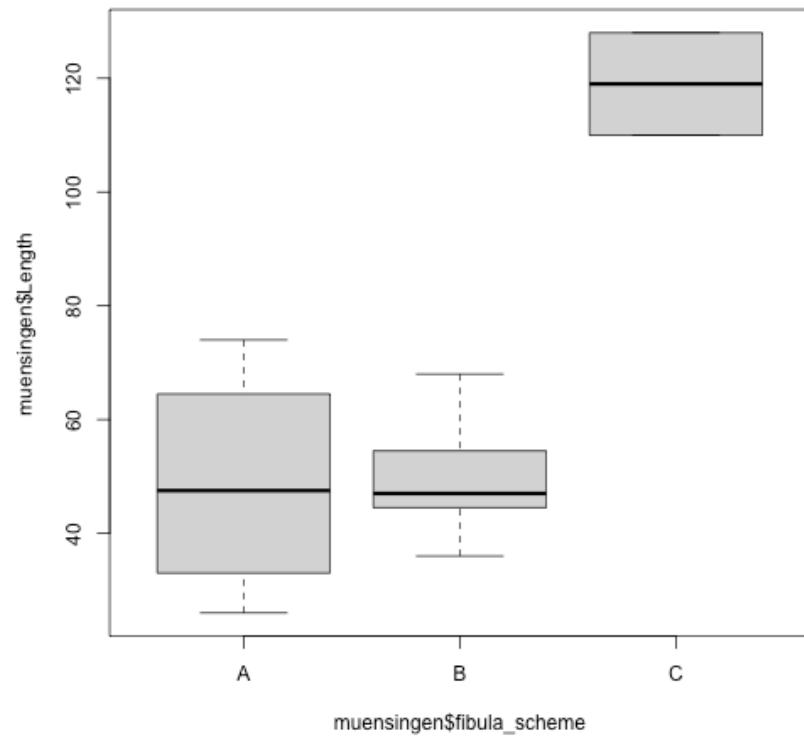
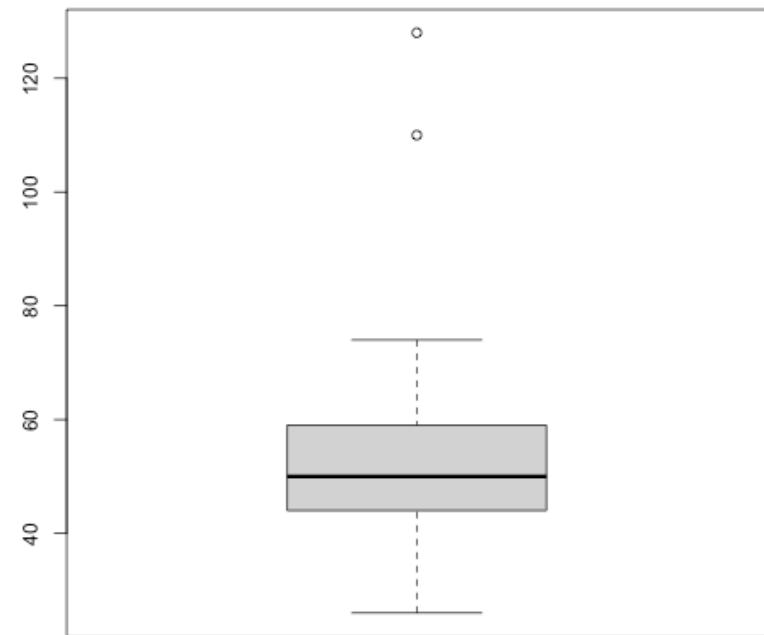
Used to display the distribution of values in a data vector of metrical (interval, ratio) scale



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

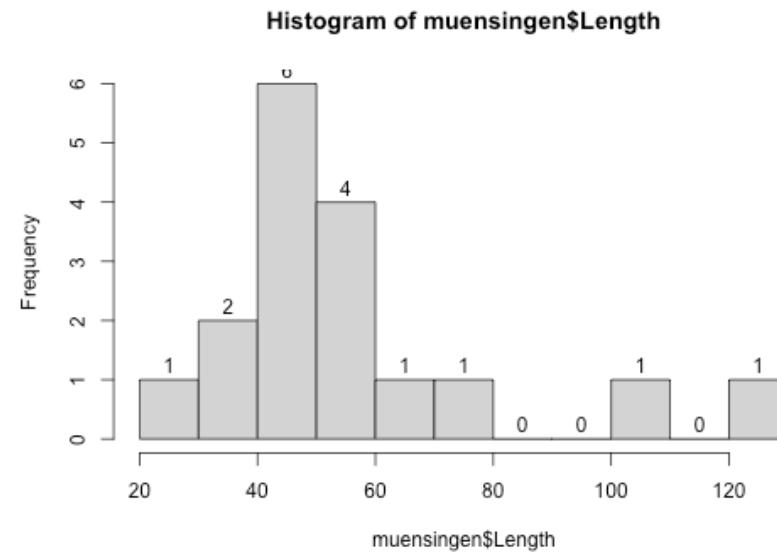
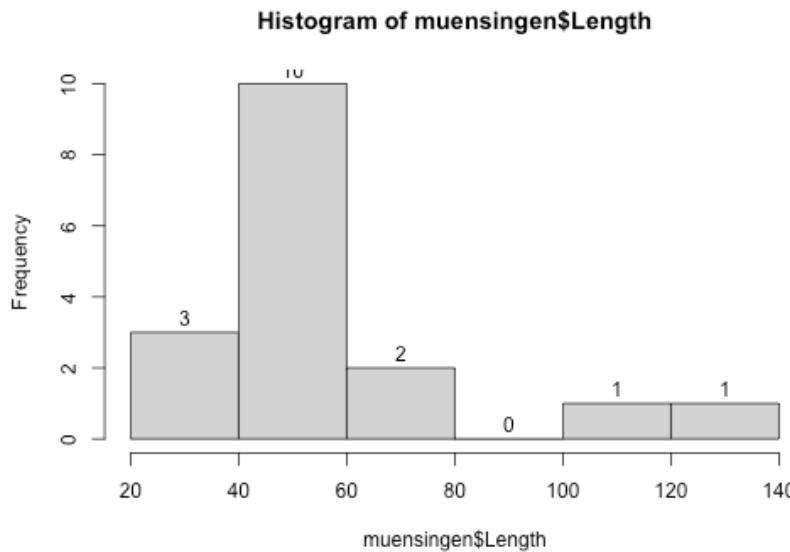


Box Plot [2]



Histogramm

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



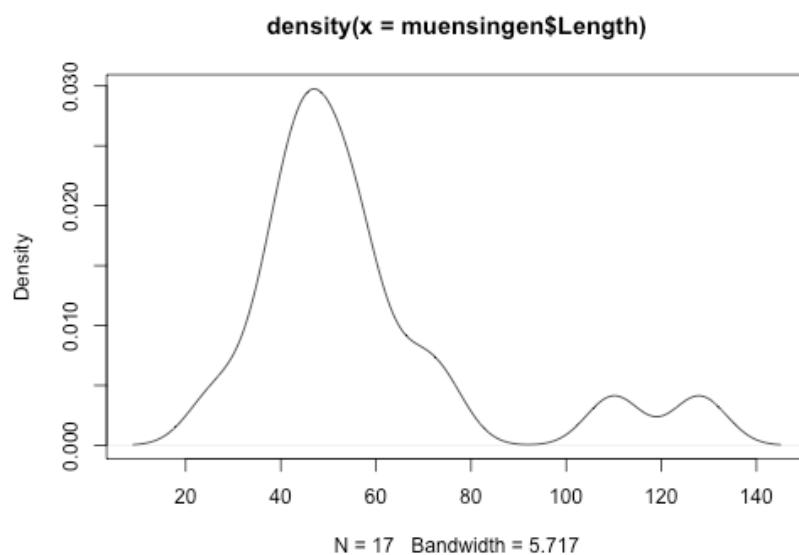
Disadvantages:

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

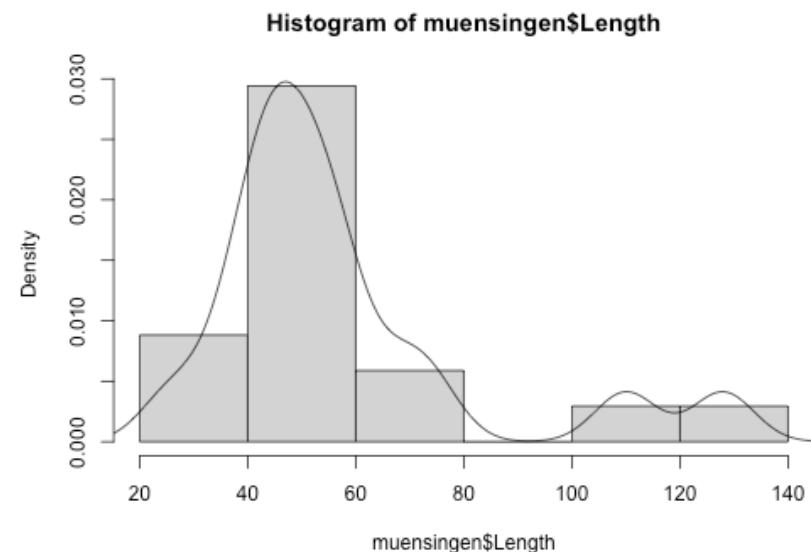
kernel smoothing (kernel density estimation)

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.



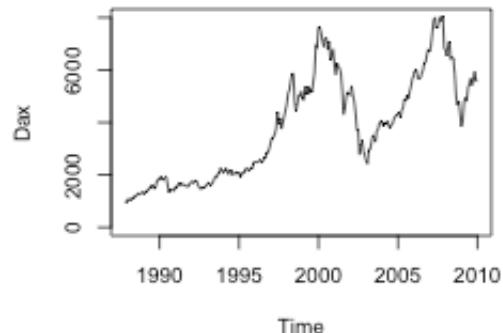
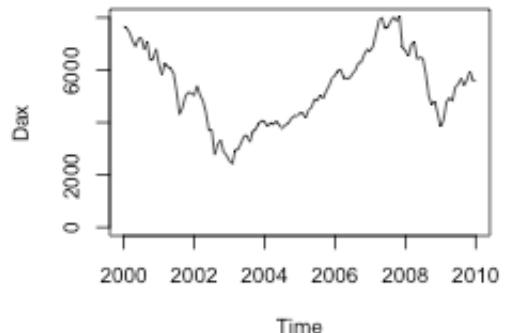
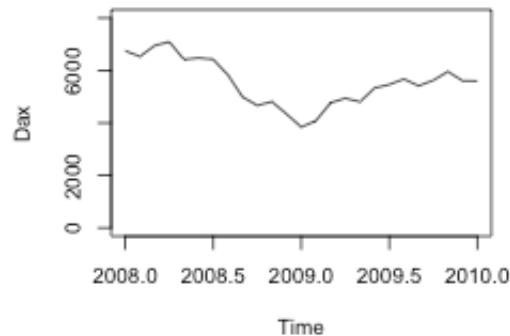
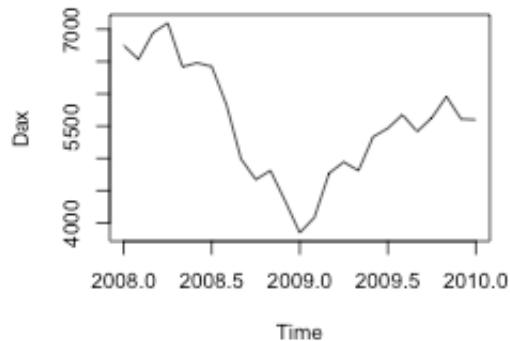
Histogram and kernel-density-plot together



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