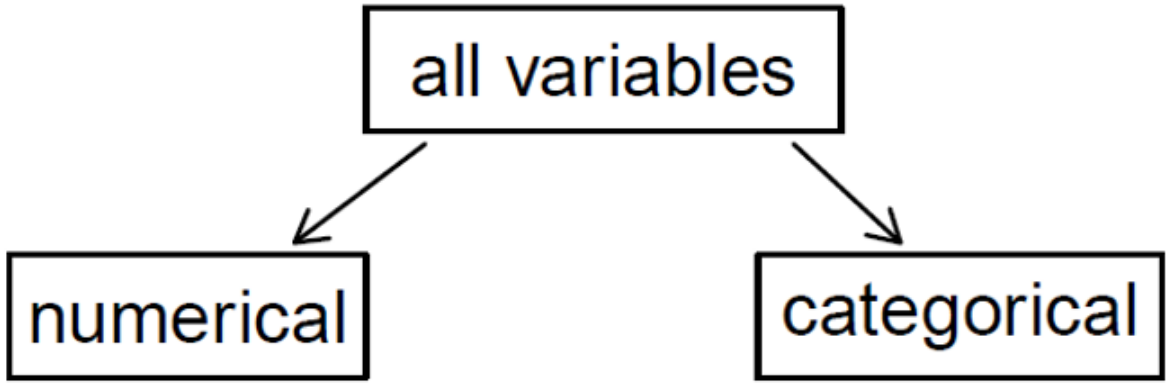
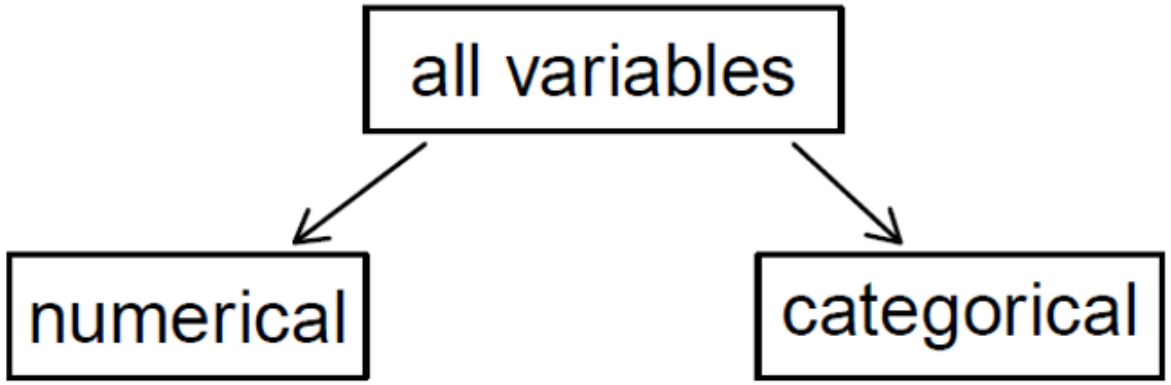
Exploratory Data Analysis

Graphical Tools

Prof. Dr. Gero Szepannek
Statistics, Business Mathematics & Machine Learning
Stralsund University of Applied Sciences

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical			
numeric			


all variables
numerical
categorical

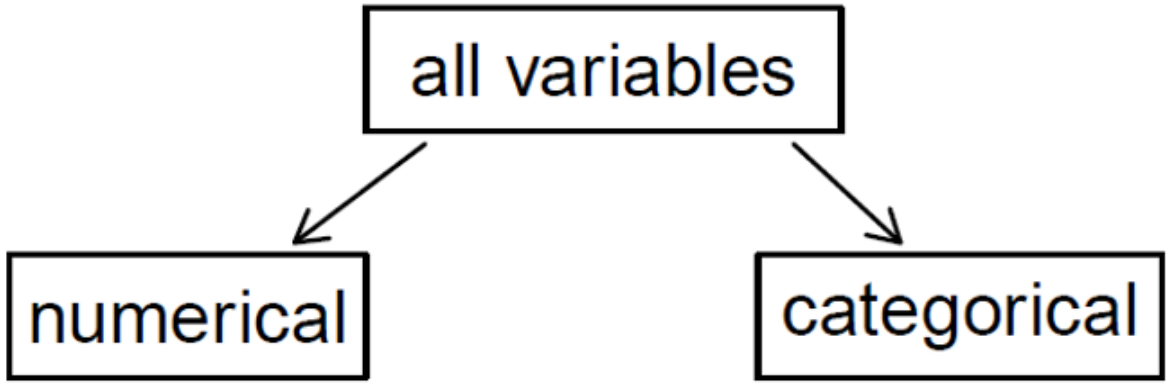


Examples:

- Size
- Number of eyes

Example:

- Colour



Examples:

- Size
- Number of eyes

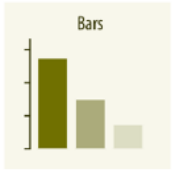
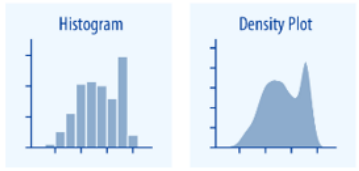
Example:

- Colour

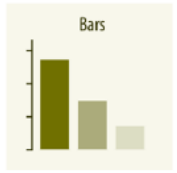
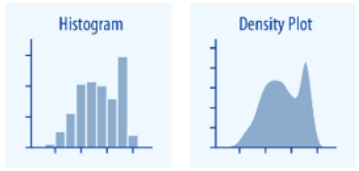
Different analysis methodology!

Variable 1	no 2nd variable!	2nd variable categorical	2nd variable numeric
categorical			
numeric			

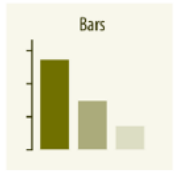
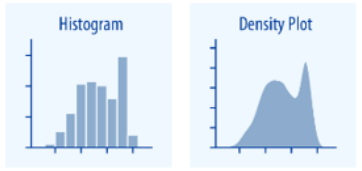
Statistical Graphics for Different Situations

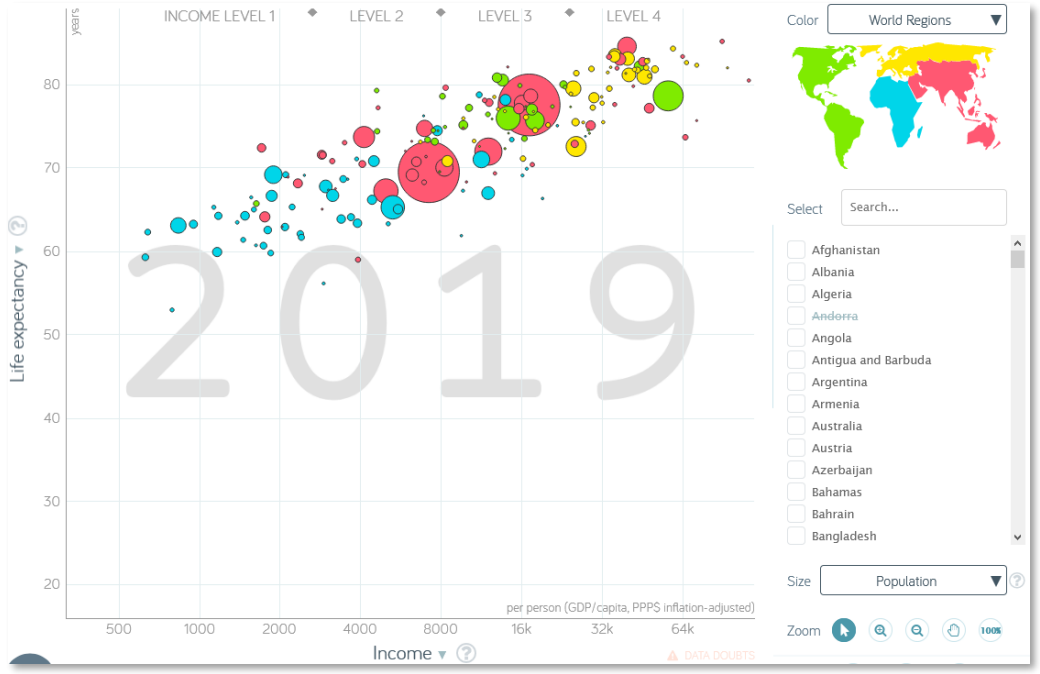
Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables		
numeric	 Histogram Kernel density plot Mean Variance / standard deviation Inter quartile range		

Statistical Graphics for Different Situations

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables		
numeric	 Histogram Kernel density plot Mean Variance / standard deviation Inter quartile range		

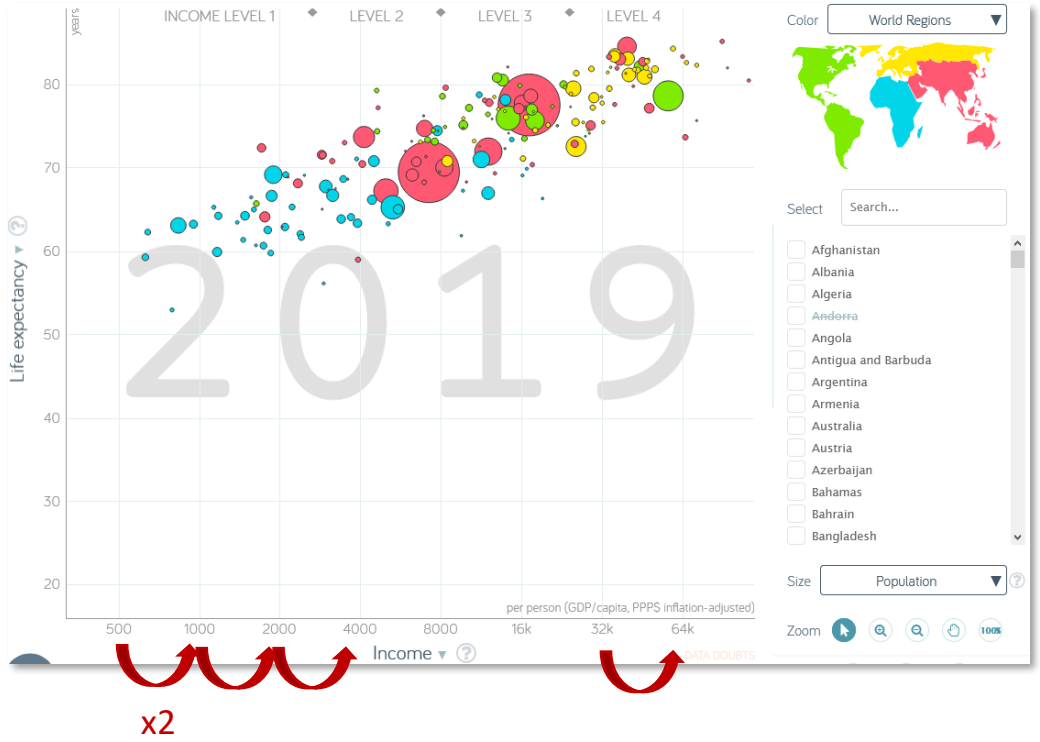
Statistical Graphics for Different Situations

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables		
numeric	 Histogram Kernel density plot Mean Variance / standard deviation Inter quartile range		 Correlation



GAPMINDER

https://www.gapminder.org/



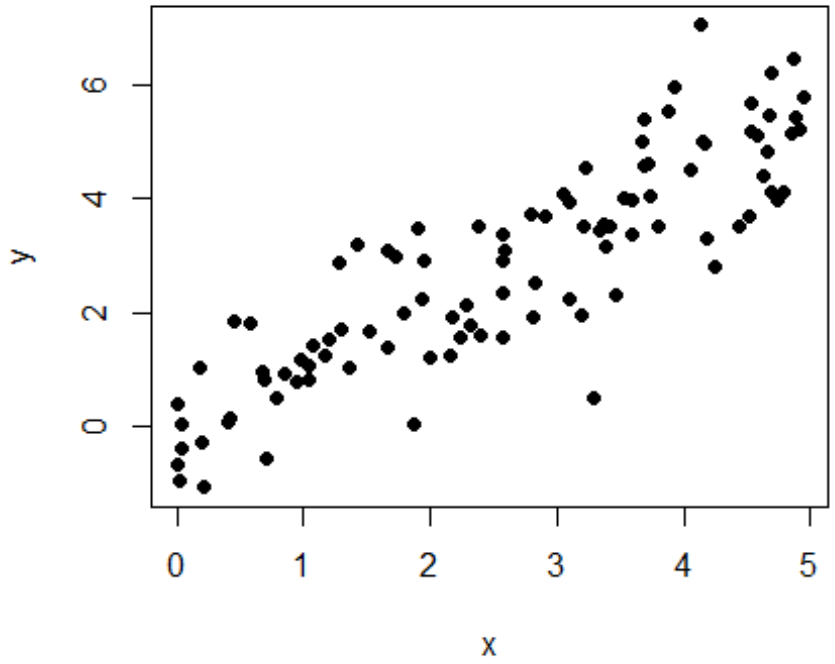
- Note: x-axis logarithmically scaled…
- This is often done if the differences are huge

A statistical measure to quantify the dependency between two numeric variables is given by the **coefficient of correlation ρ** :

Interpretation:

$-1 \leq \rho \leq 1$ where:

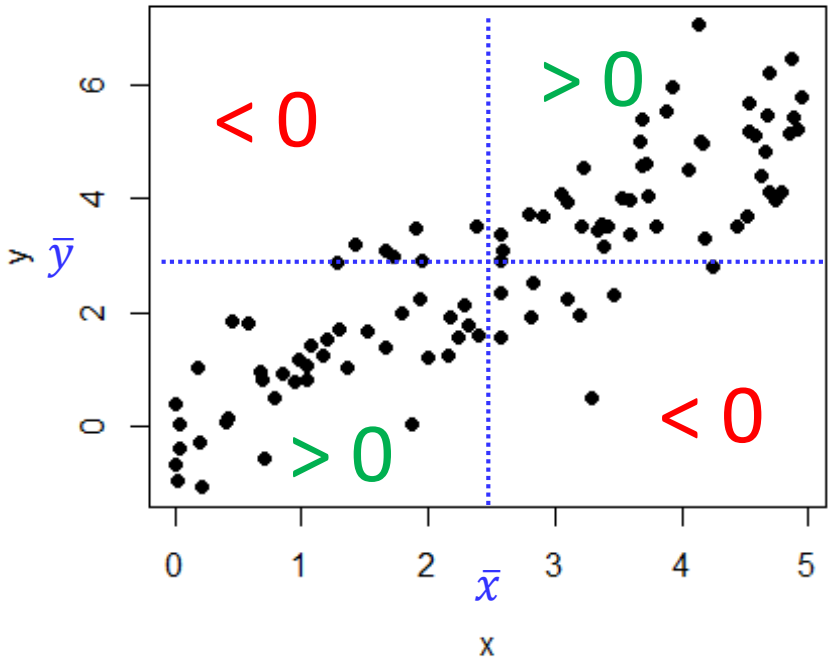
Correlation	Interpretation
ρ > 0	Positive dependency
ρ = 0	No (linear) dependency
ρ < 0	Negative dependency



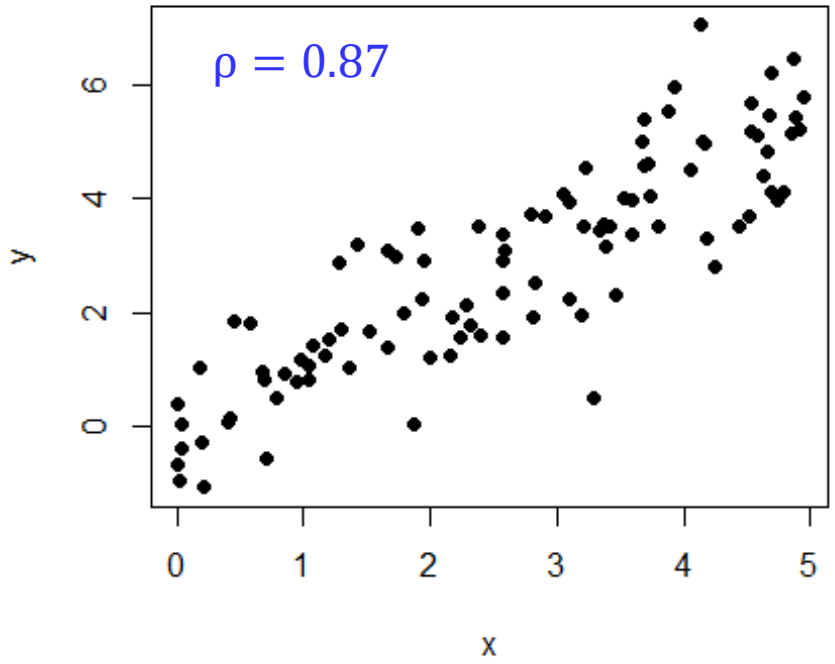
…comparison with the mean of x and y

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_j - \bar{y})^2}}$$

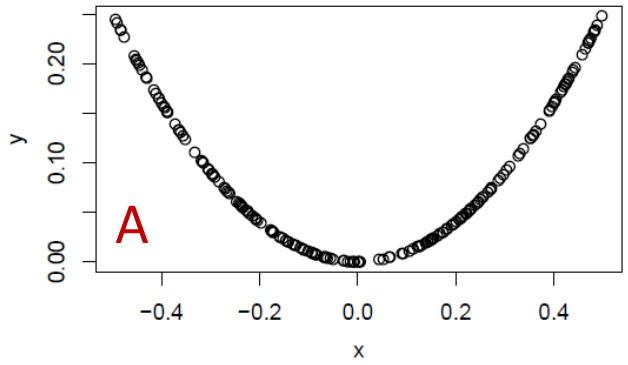
…only for scaling $-1 \leq \rho \leq 1$

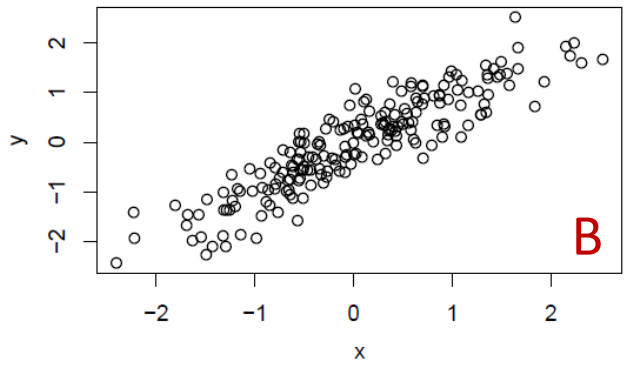


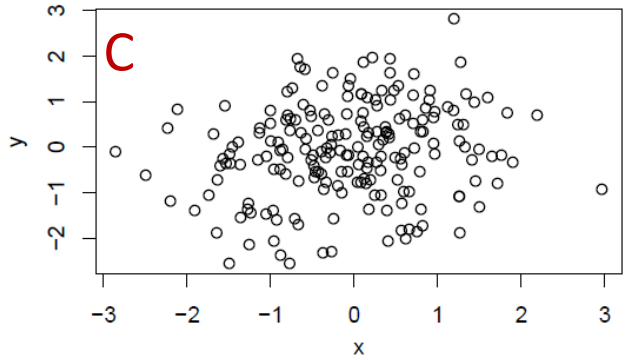
Correlation	Interpretation
$\rho > 0$	Positive dependency
$\rho = 0$	No (linear) dependency
$\rho < 0$	Negative dependency

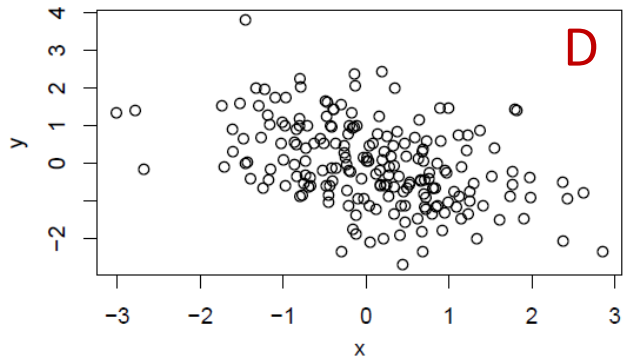


But note: No correlation (i.e. $\rho = 0$) $\not\Rightarrow$ **independence (!)** (just: no <u>linear</u> dependency).



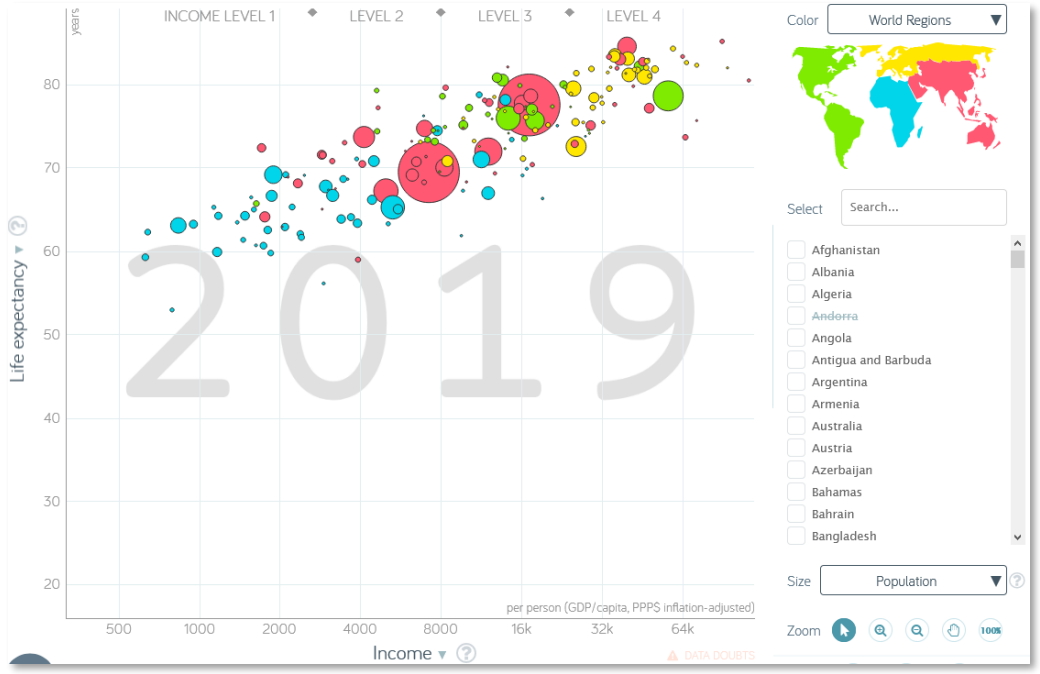


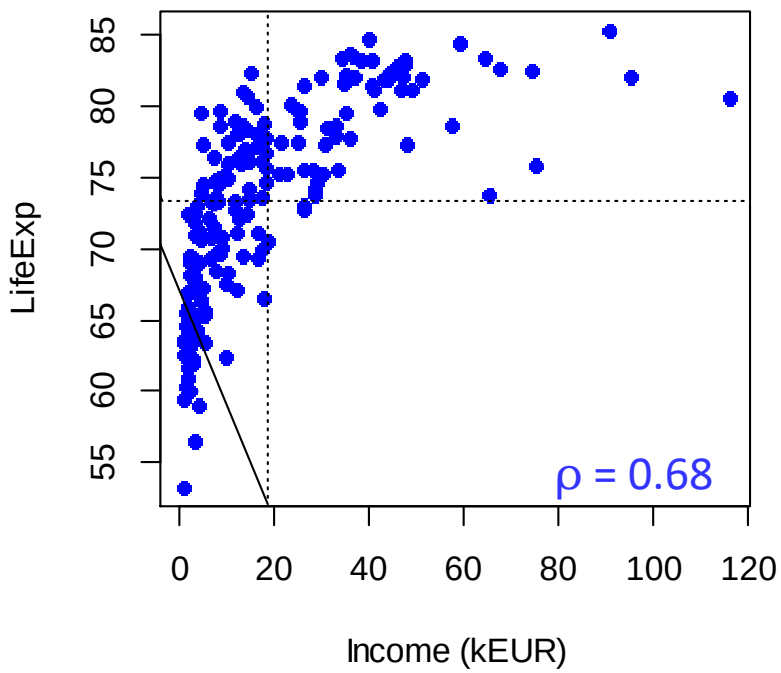




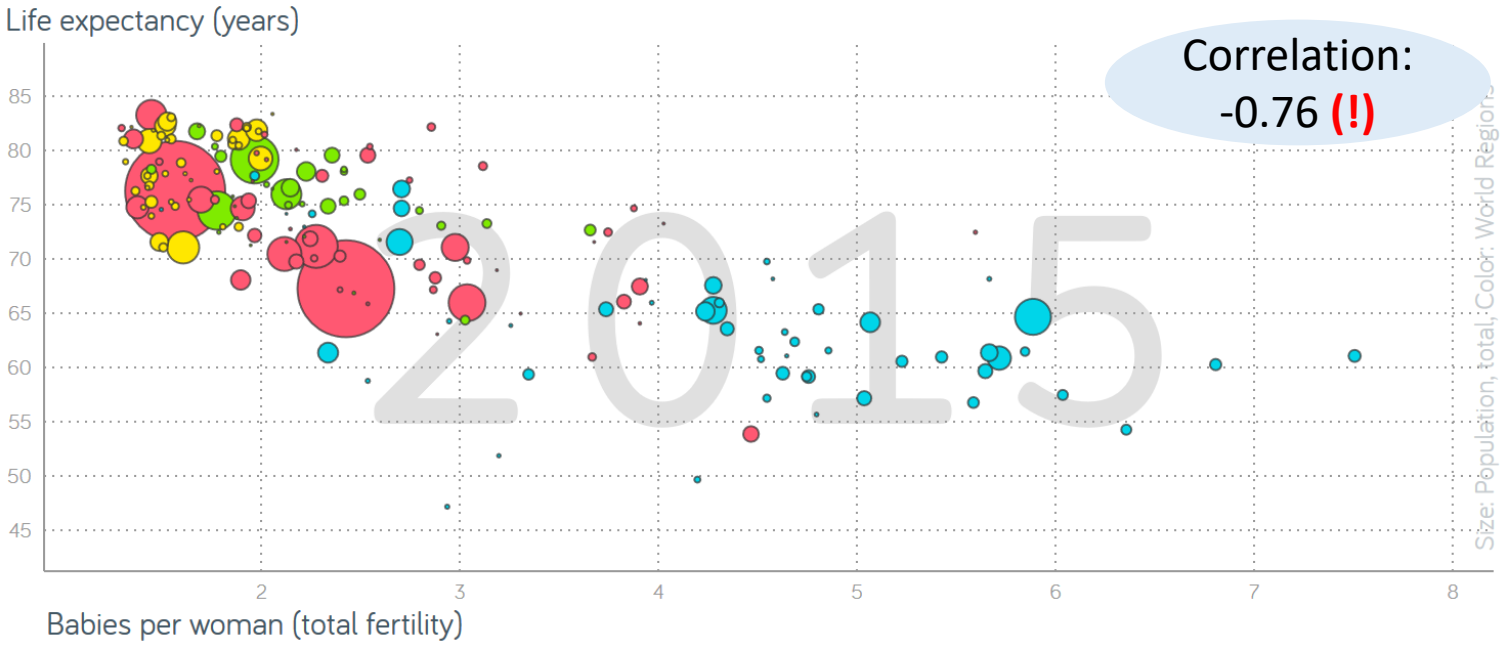
Which is the corresponding plot to a correlation of:

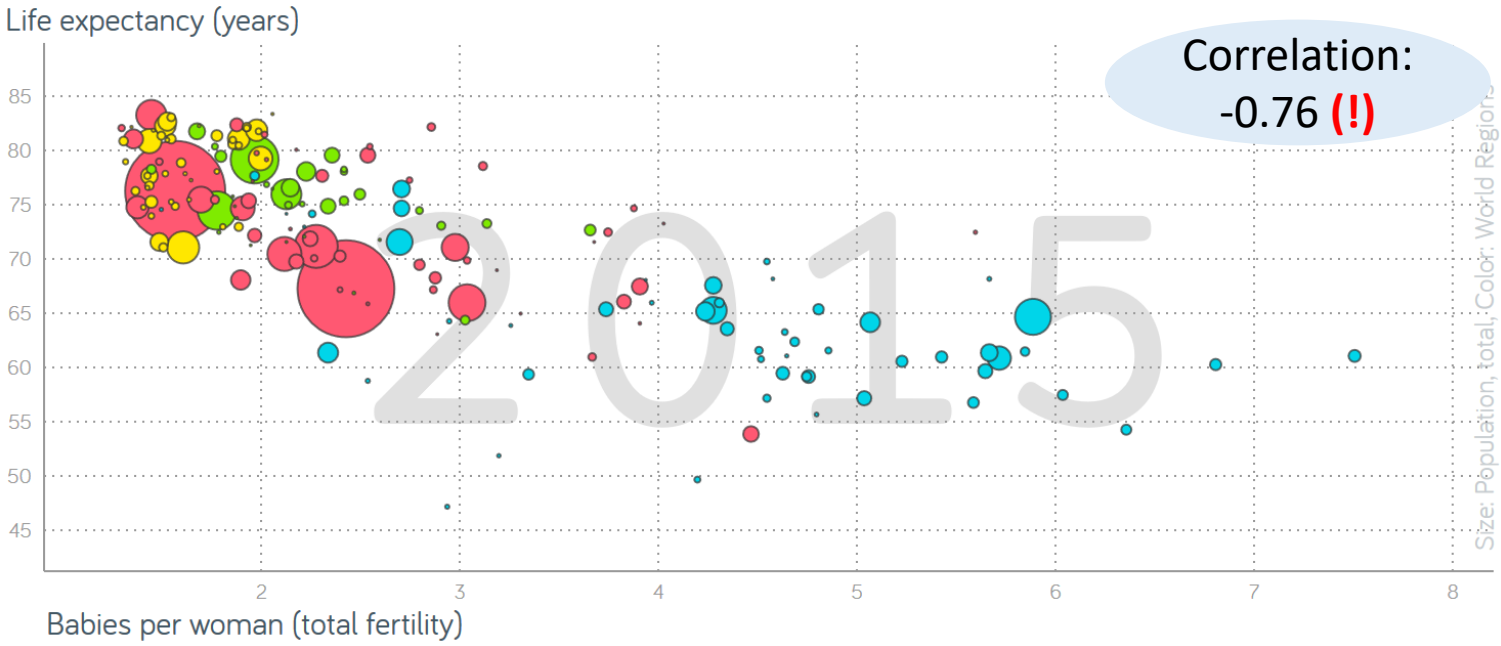
- ❑ **0.9**
- ❑ **-0.42**





ρ = 0.68 → Strong positive dependency btw income and life expectancy!

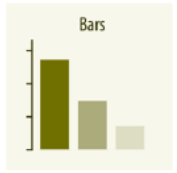
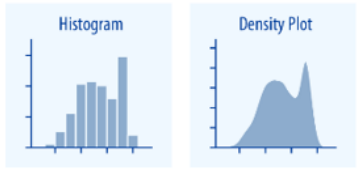

Life expectancy (years)
85
80
75
70
65
60
55
50
45
2015
2
3
4
5
6
7
8
Babies per woman (total fertility)
Size: Population, total; Color: World Regions
Correlation:
-0.76 (!)



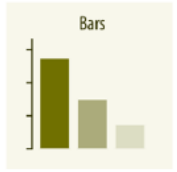
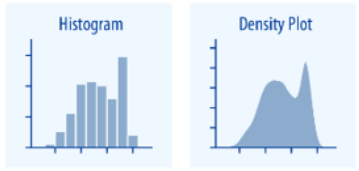

Why We Should Teach Causal Inference: Examples in Linear Regression With Simulated Data

(Lübke, K., Gehrke, M., Horst J. and Szepannek, G., 2020)

Statistical Graphics for Different Situations

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables		
numeric	 Histogram Kernel density plot mean Variance / standard deviation Inter quartile range		 Correlation

Variable 1

	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables	 Contingency tables χ^2 Test Cramer's V	
numeric	 Histogram Kernel density plot mean Variance / standard deviation Inter quartile range		 Correlation

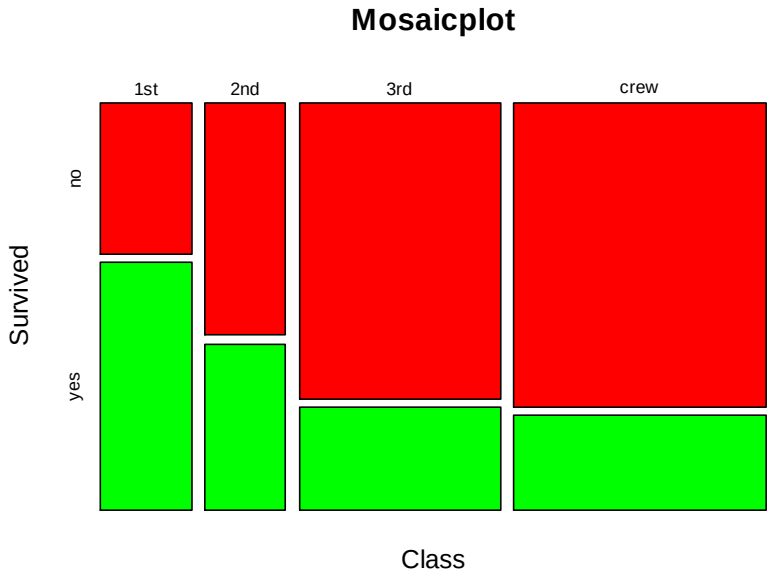
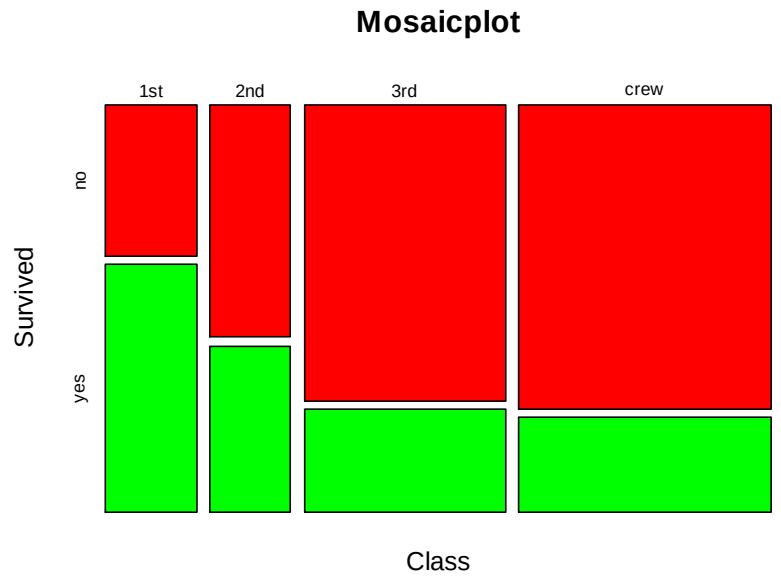


Figure taken from: https://www.geo.de/geolino/mensch/10493-rtkl-geschichte-die-letzte-nacht-auf-der-titanic

$P(survived|class)$



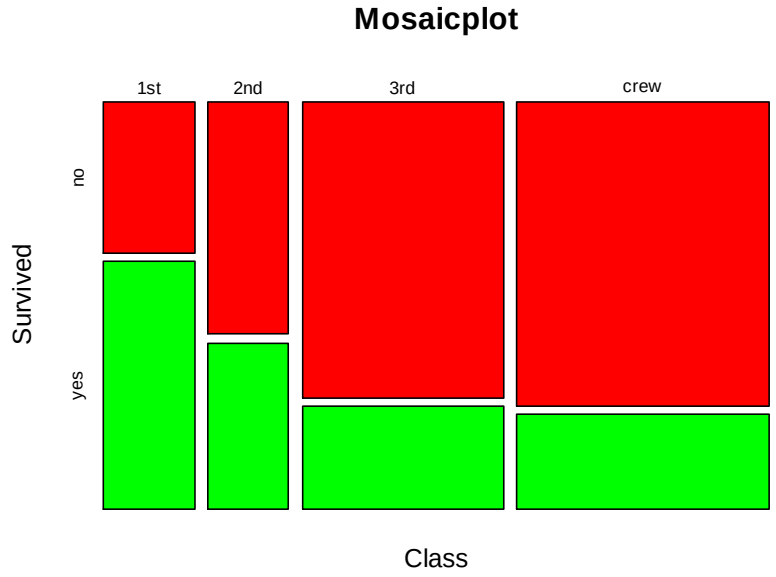
$P(class)$

Absolute frequencies

	1st	2nd	3rd	crew	total
no	123	166	528	679	1496
yes	201	118	181	211	711
total	324	284	709	890	2207

Conditional frequencies

	1st	2nd	3rd	crew	total
no	0,37963	0,58451	0,74471	0,76292	1496
yes	0,62037	0,41549	0,25529	0,23708	711
total	324	284	709	890	2207



observed

			Σ
Frau	50	20	
Mann	10	20	
Σ			

observed

			Σ
Frau	50	20	
Mann	10	20	
Σ			

expected

			Σ
Frau			
Mann			
Σ			

What counts could we expect if gender and preference were independent?

observed

			Σ
Frau	50	20	
Mann	10	20	
Σ			

$$\left(\text{observed} - \text{expected} \right)^2$$

expected

			Σ
Frau			
Mann			
Σ			

$$\frac{\left(\text{observed} - \text{expected} \right)^2}{\text{expected}}$$

expected

			Σ
Frau			
Mann			
Σ			

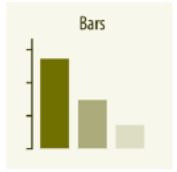
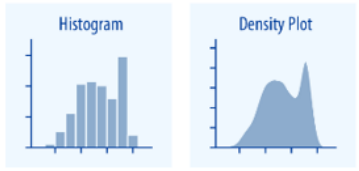
$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

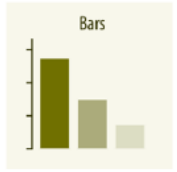
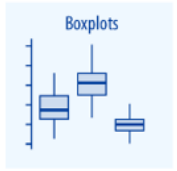
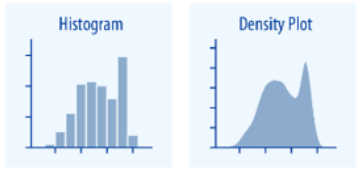
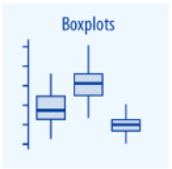
$$V = \sqrt{\frac{\chi^2/n}{\min(c-1, r-1)}}$$

n: #observations
c/r: #columns/rows of the table

$0 \leq V \leq 1$ measures the dependency between two categorical variables.

Statistical Graphics for Different Situations

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables	 Contingency tables χ^2 Test Cramer's V	
numeric	 Histogram Kernel density plot mean Variance / standard deviation Inter quartile range		 Correlation

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables	 Contingency tables χ^2 Test Cramer's V	 Boxplots Mean | group Variance | group
numeric	 Histogram Kernel density plot mean Variance / standard deviation Inter quartile range	 Boxplots Mean | group Variance | group	 Correlation

- Visualization of the distribution based on five key measures: min, max, $q_{0.75}$, $q_{0.25}$ & median.
- The box covers 50% of the data.
- ‘Whiskers’: $q_{0.75}/q_{0.25} \pm {}^{3}/_{2}\, IQR$ or the maximum/minimum.
- …Points outside the whiskers are suspicious to be outliers.

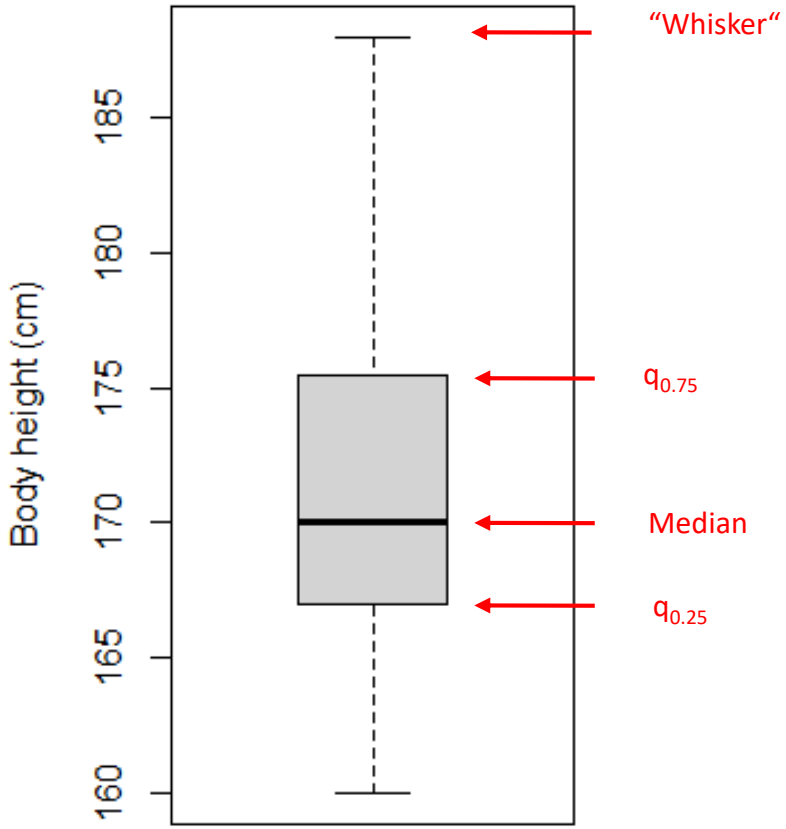
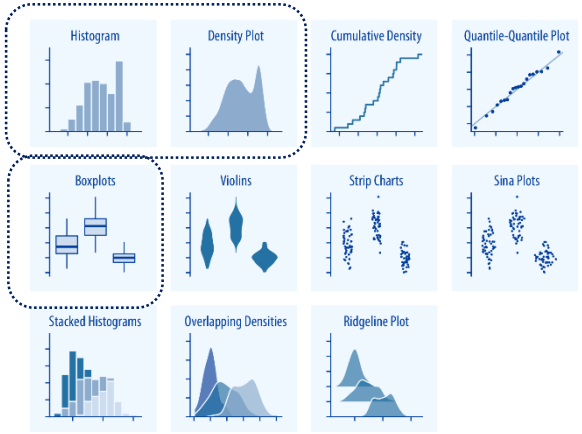


Figure taken from Diez, Cetinkaya-Rundel and Barr (2019): OpenIntro Statistics. https://www.openintro.org/book/os/

Overview Data Visualzation Techniques

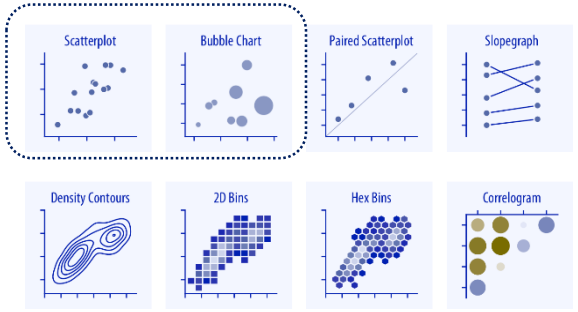
Amounts

Distributions

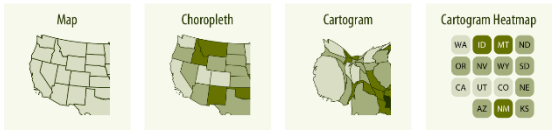


Proportions

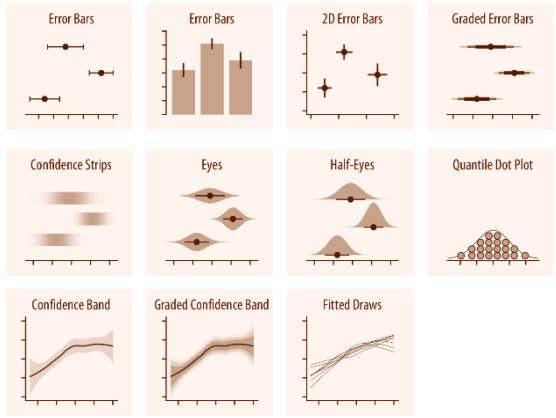
Dependency

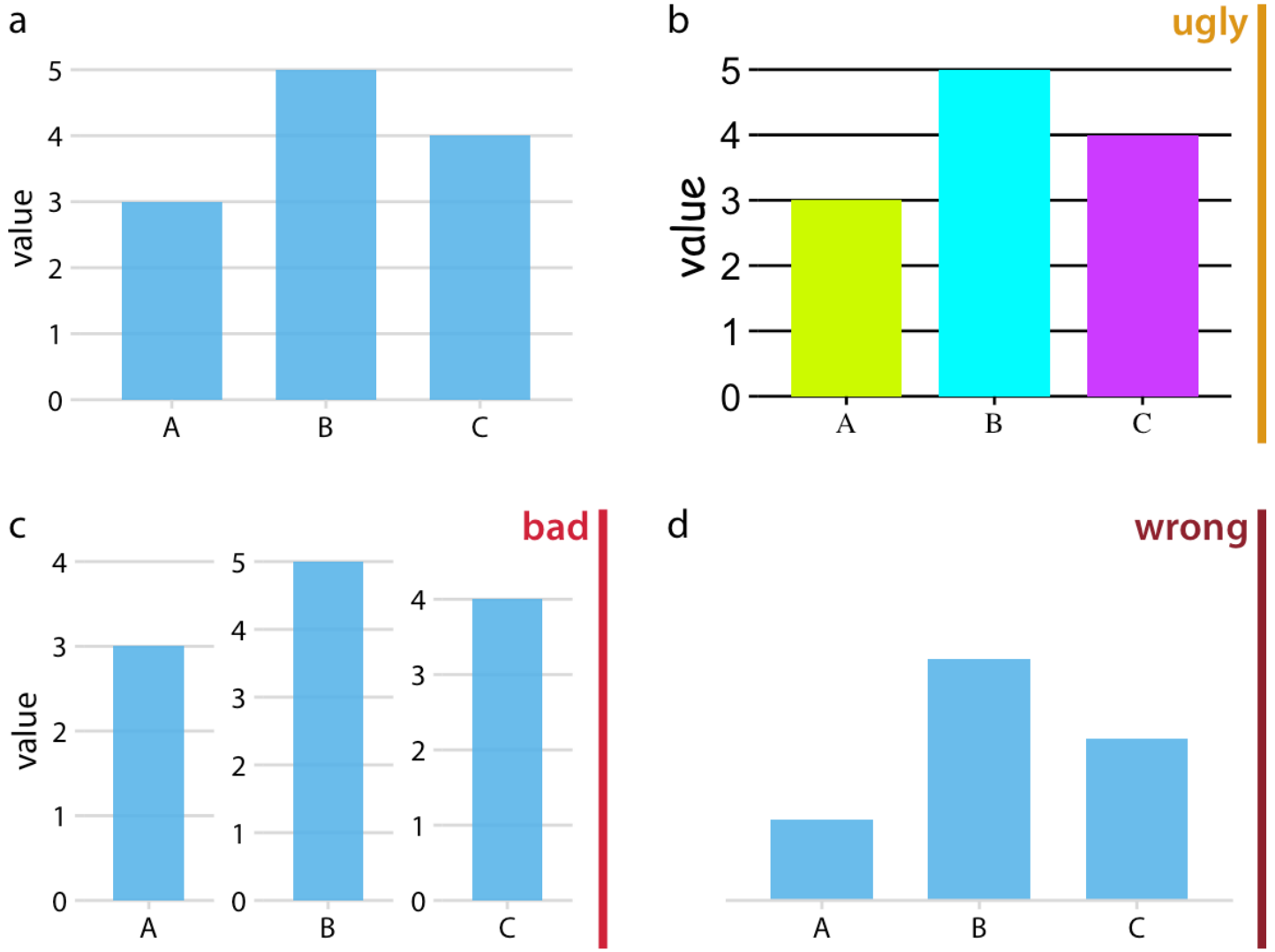


Geodata



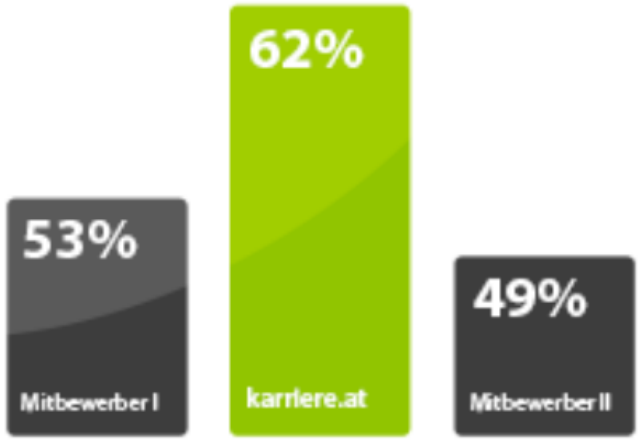
Uncertainty




a
value
5
4
3
2
1
0
A
B
C
b
ugly
value
5
4
3
2
1
0
A
B
C
c
bad
value
4
3
2
1
0
A
5
4
3
2
1
0
B
4
3
2
1
0
C
d
wrong
A
B
C

Höchste Bekanntheit

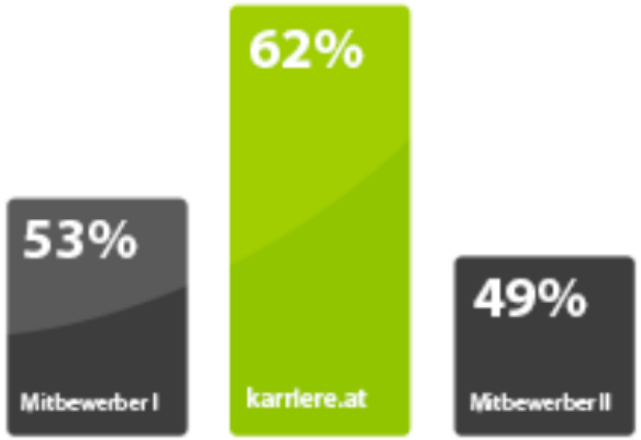
Fast 2/3 der Arbeitnehmer kennen karriere.at.
Im Mitbewerbsvergleich ist das spitze.



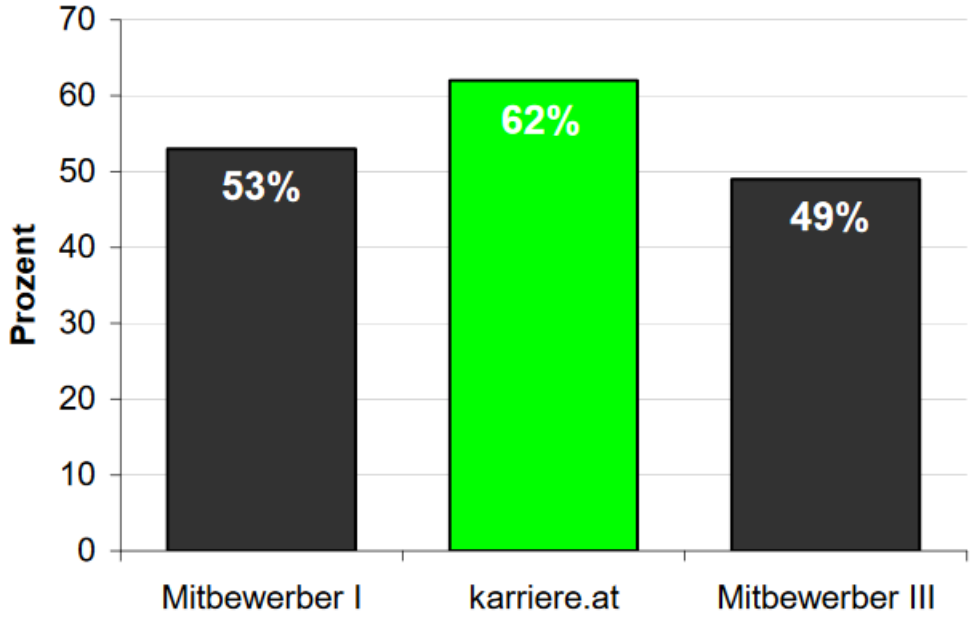
(gefunden am 12. November 2014 auf http://www.karriere.at/hr)

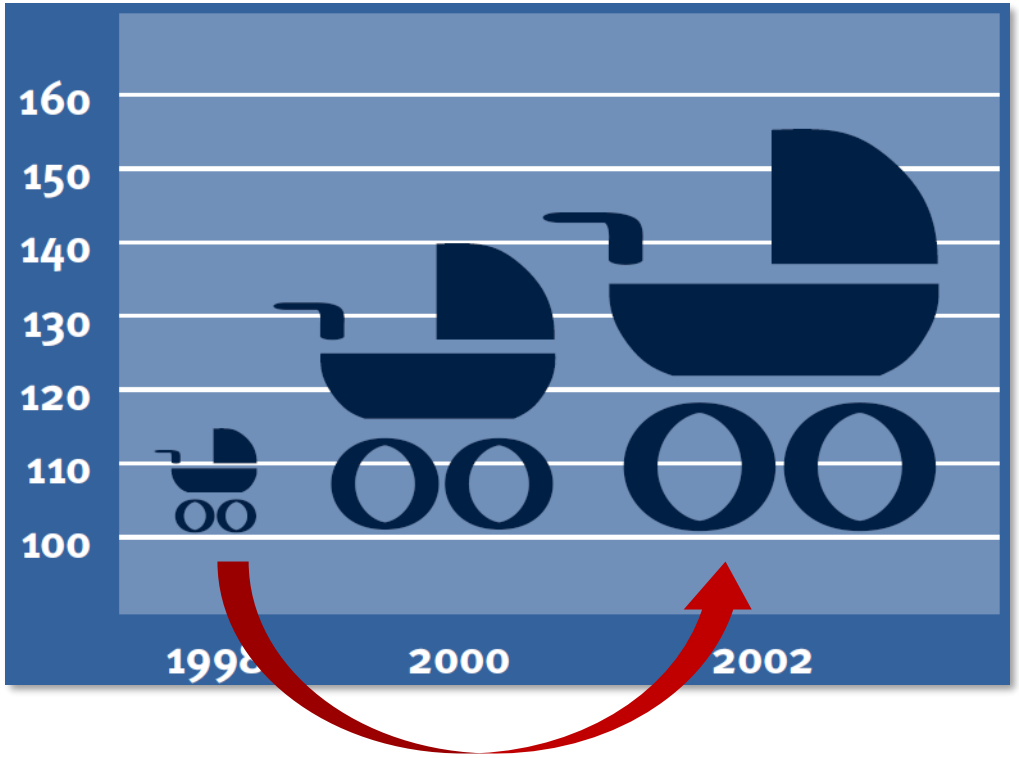
Höchste Bekanntheit

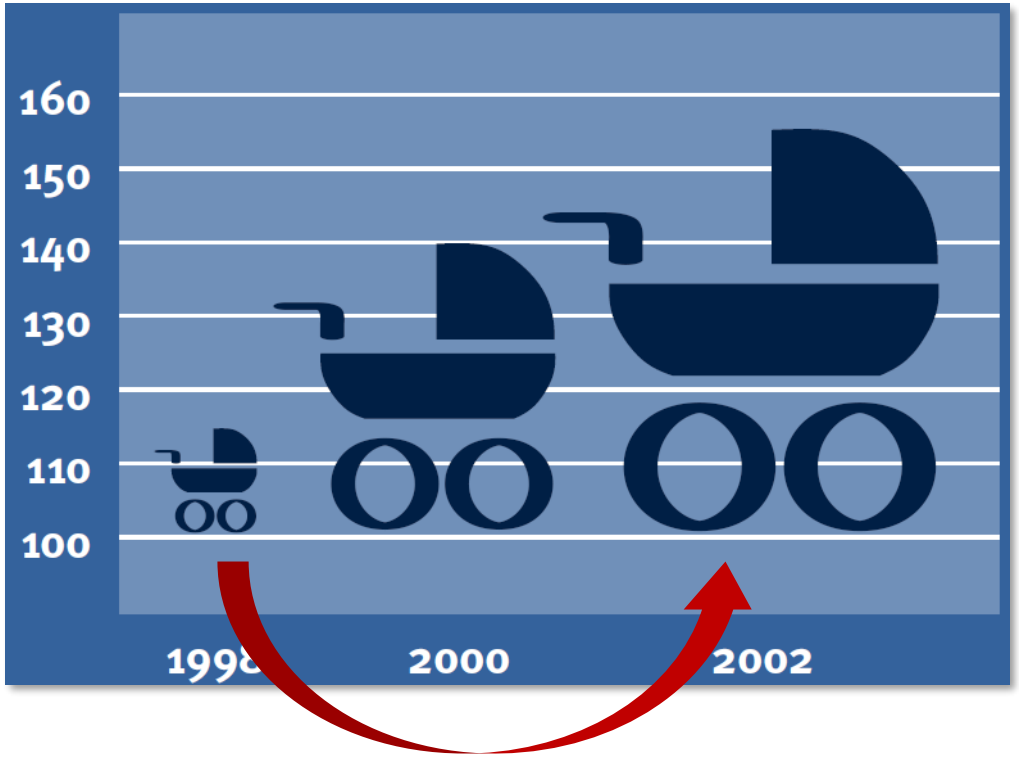
Fast 2/3 der Arbeitnehmer kennen karriere.at.
Im Mitbewerbsvergleich ist das spitze.

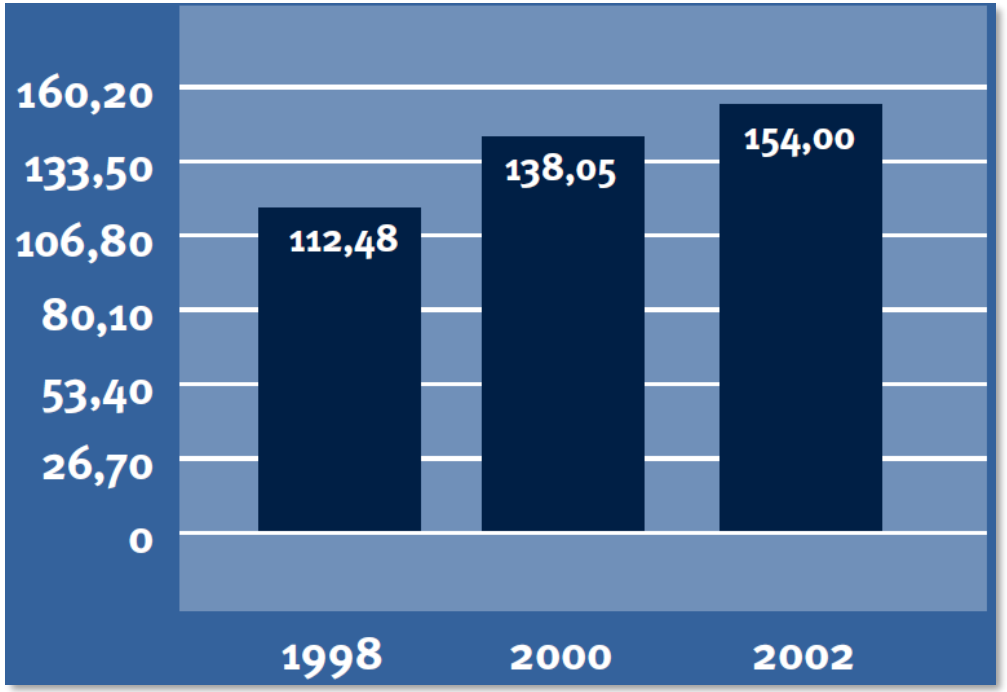


(gefunden am 12. November 2014 auf http://www.karriere.at/hr)



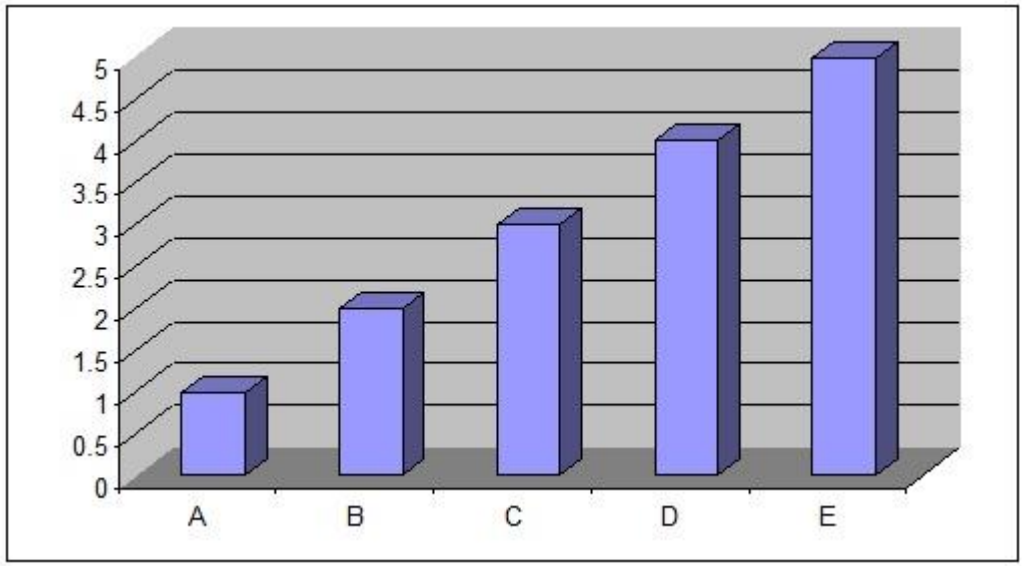

160
150
140
130
120
110
100
2000
2002



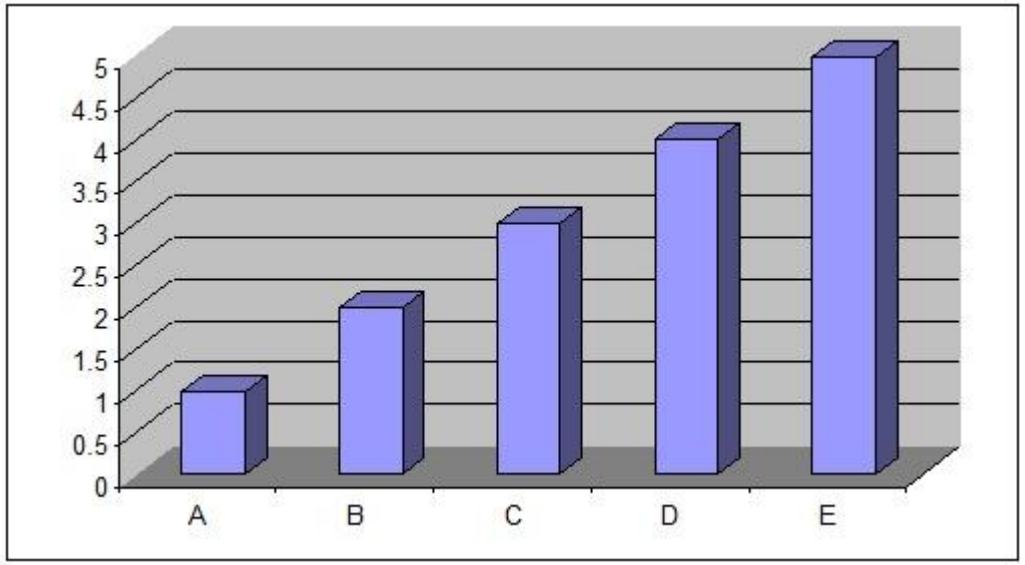


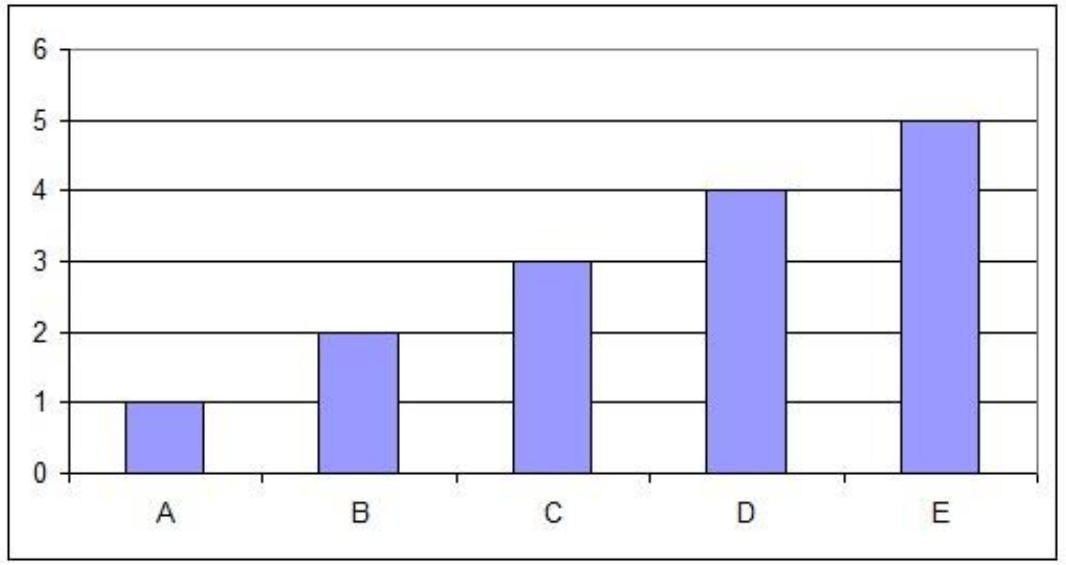
Source: http://www.wdr.de/tv/applications/fernsehen/wissen/quarks/pdf/Q_Zahlen.pdf

…bad!



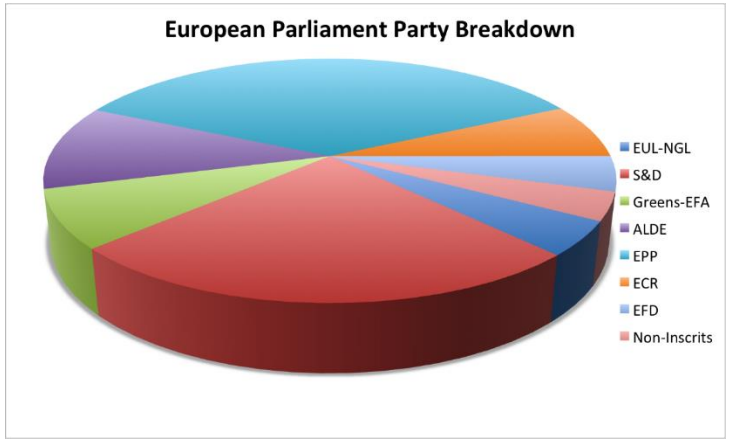
Source: http://consultantjournal.com/blog/use-3d-charts-at-your-own-risk



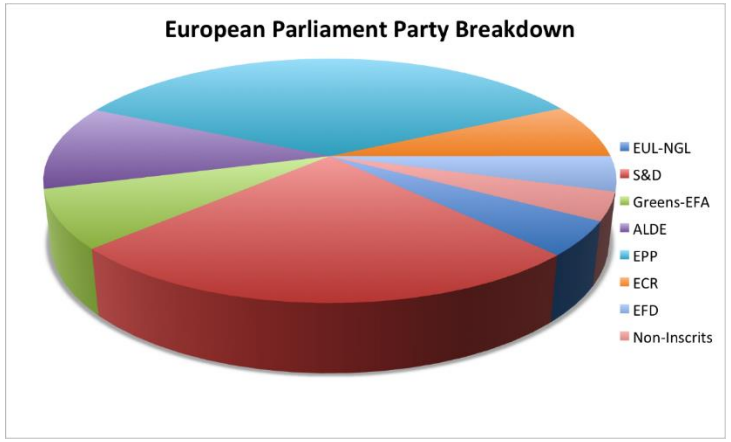


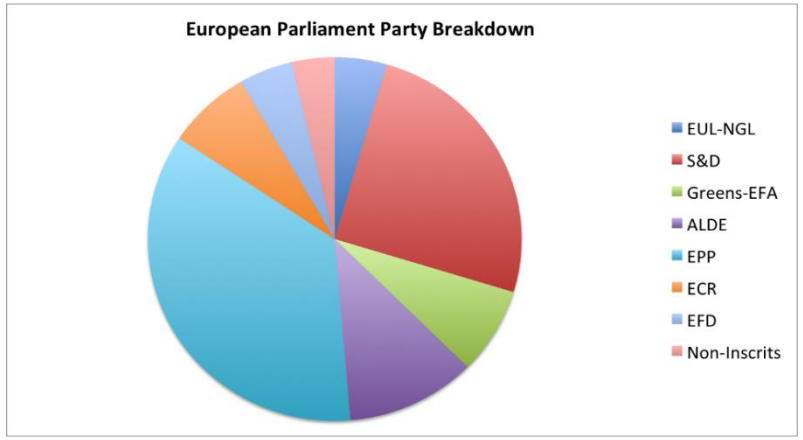
Source: http://consultantjournal.com/blog/use-3d-charts-at-your-own-risk

…bad!

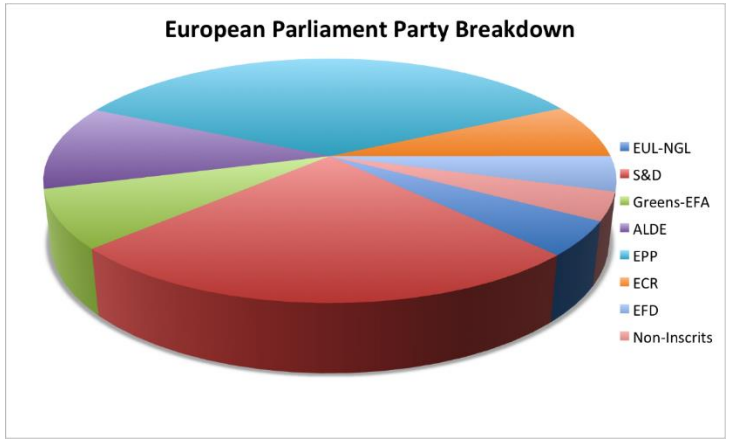


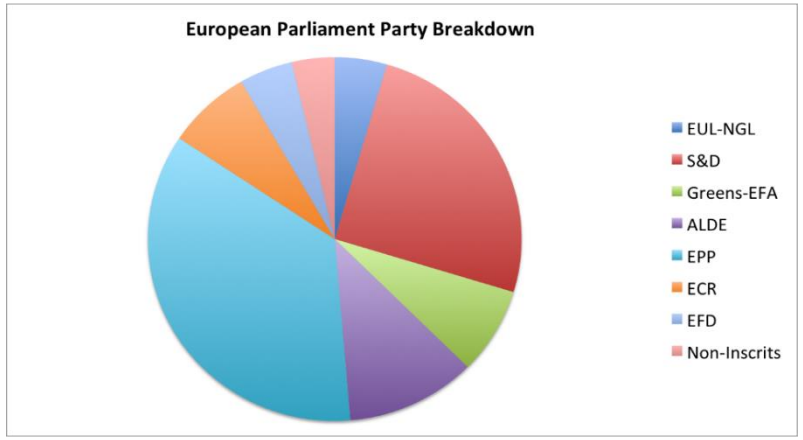
Source: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T

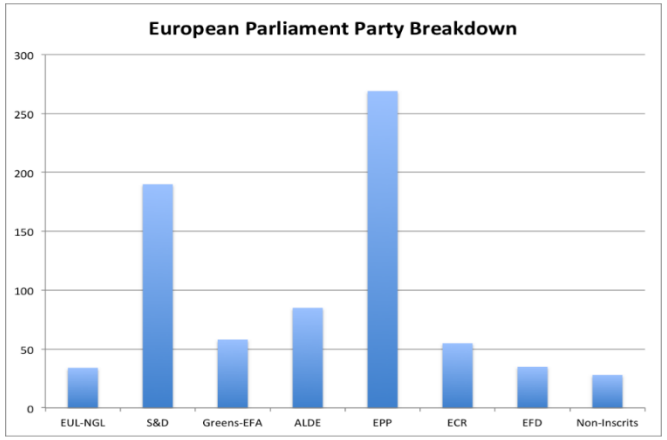




Source: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T

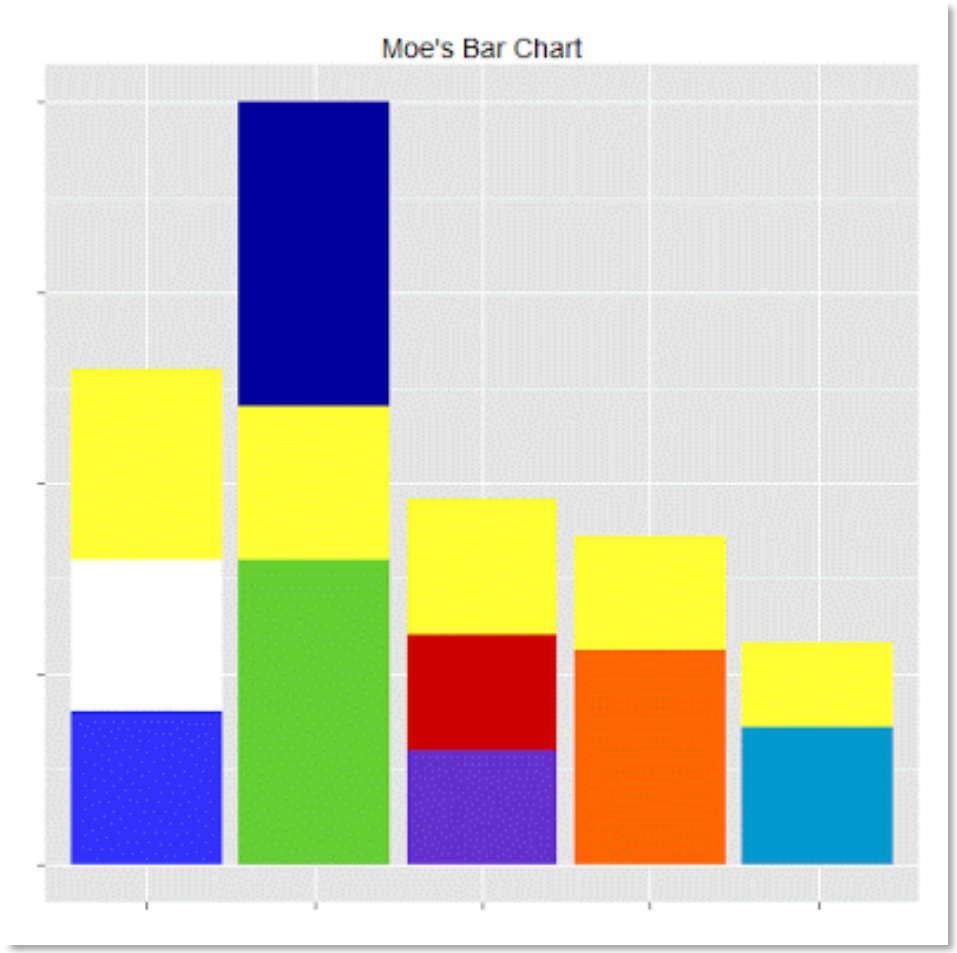


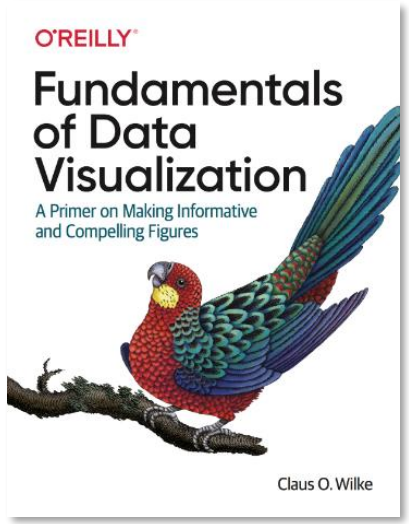




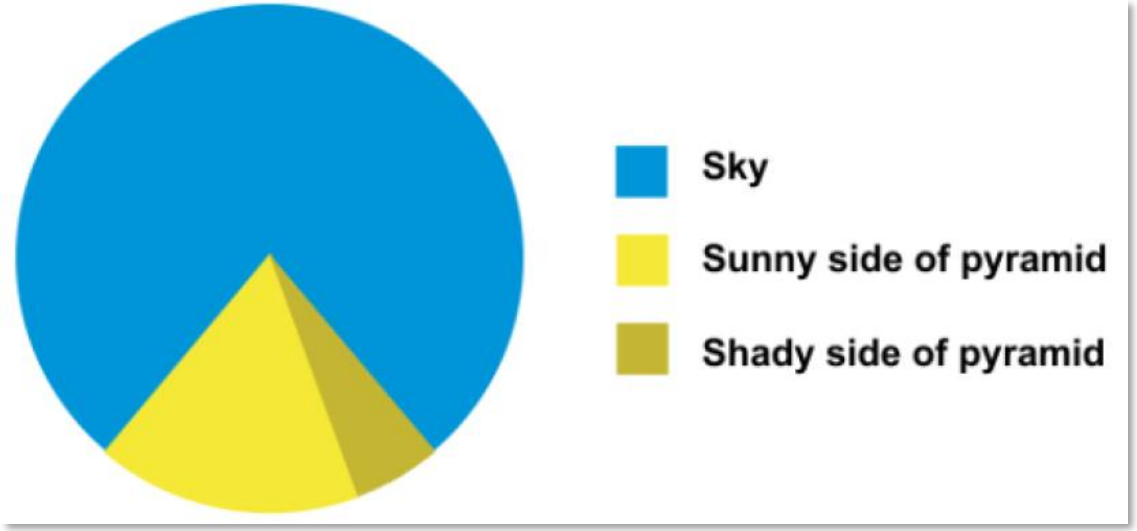
Source: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T

Source: A. Quatember: *Statistischer Unsinn – Wenn Medien an der Prozenthürde scheitern*, Springer Spektrum.


Moe's Bar Chart


O'REILLY
Fundamentals of Data Visualization
A Primer on Making Informative and Compelling Figures
Claus O. Wilke


Andreas Quatember
Statistischer Unsinn
Wenn Medien an der Prozenthürde scheitern
Springer Spektrum


Sky
Sunny side of pyramid
Shady side of pyramid