Data Analysis

Analyzing Dependency

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

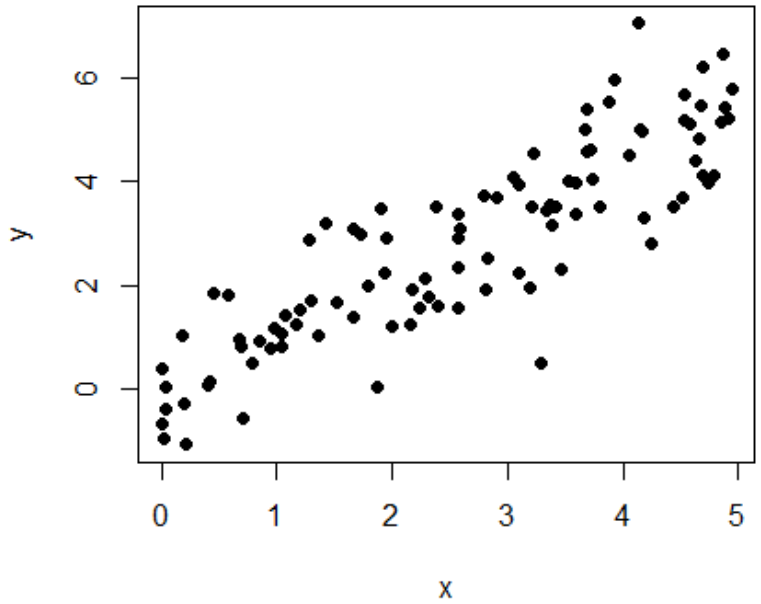
Quantifying Dependency

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

Interpretation:

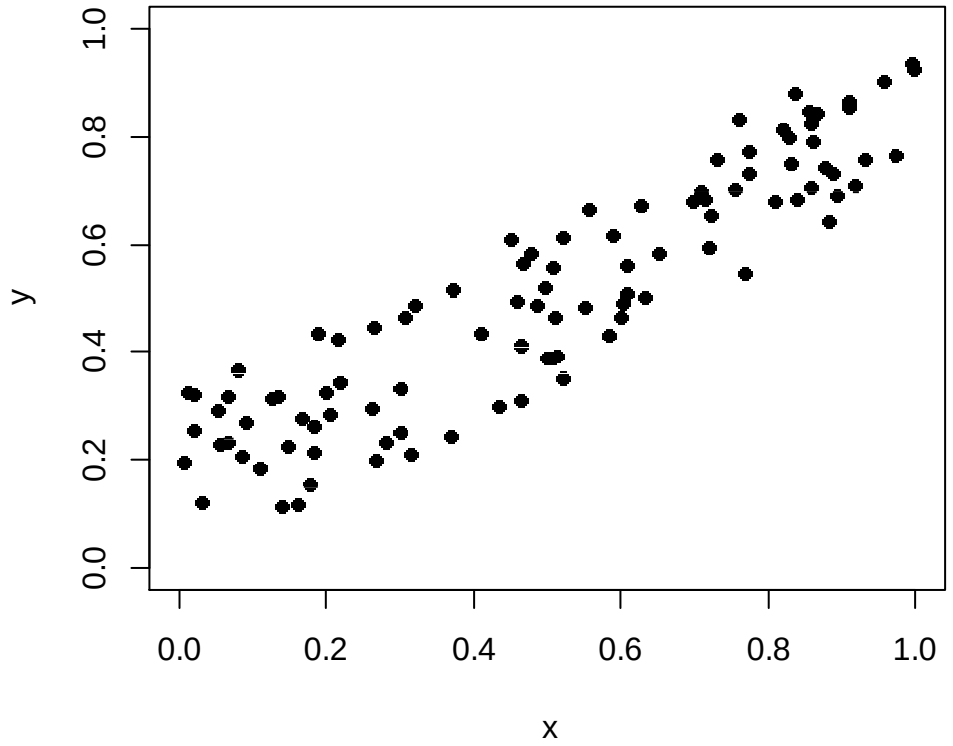
$-1 \leq \rho \leq 1$ where:

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



Understanding Correlation…

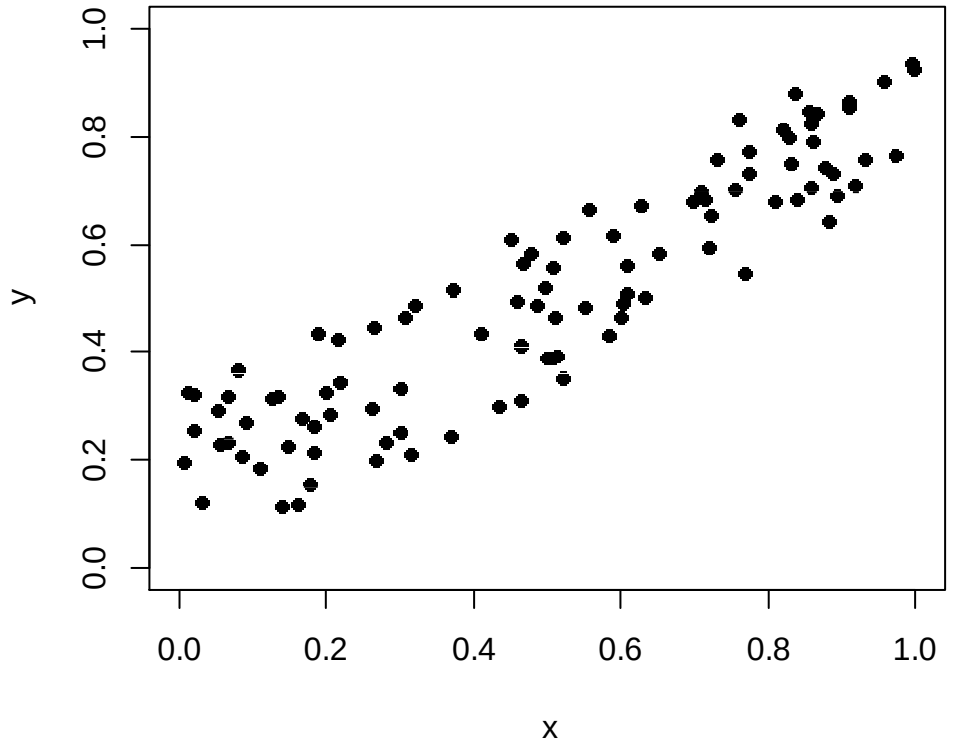
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_j - \bar{y})^2}}$$



Understanding Correlation…

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

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

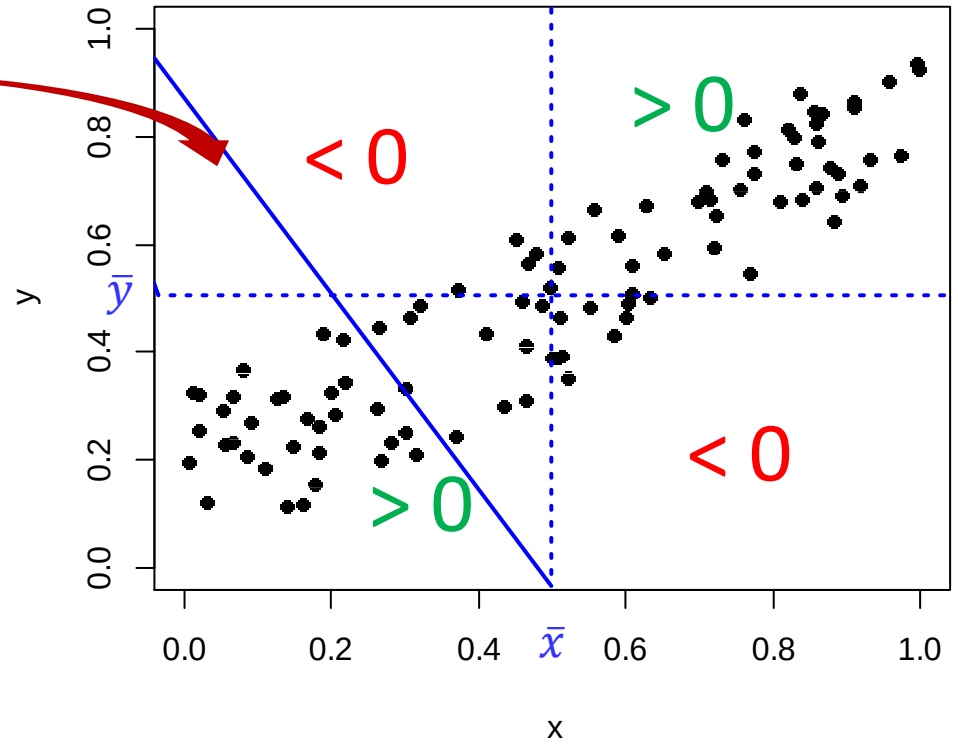


Understanding Correlation

…comparison with the mean of x and y

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

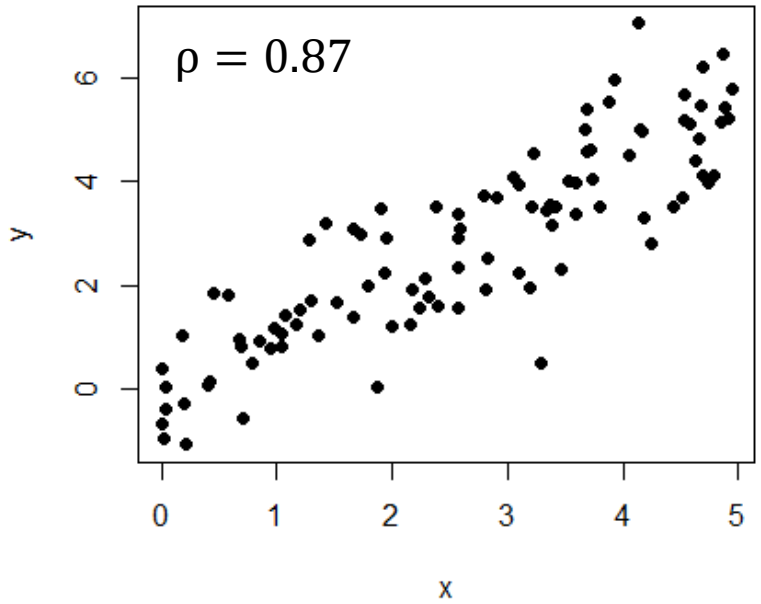
…only for scaling $-1 \leq \rho \leq 1$



Interpretation:

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

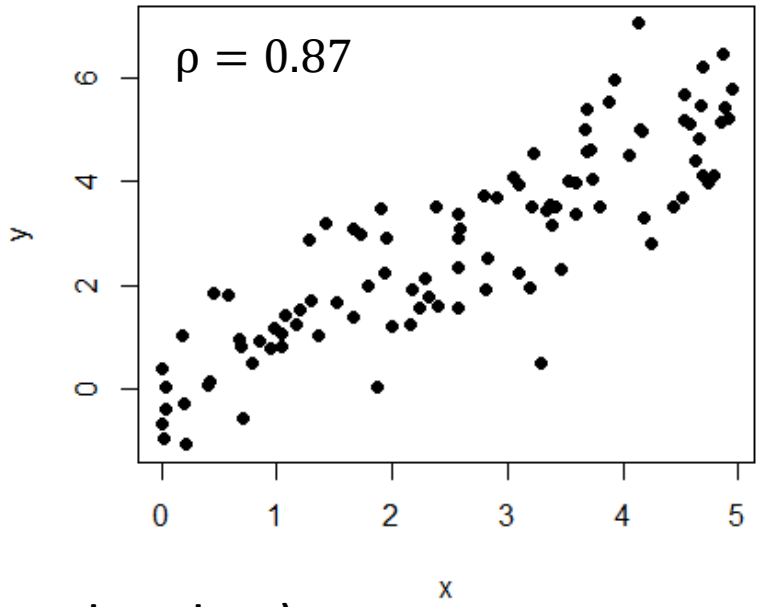
Correlation	Interpretation
ρ > 0	Positive dependency
ρ = 0	No (linear) dependency
ρ < 0	Negative dependency



Interpretation:

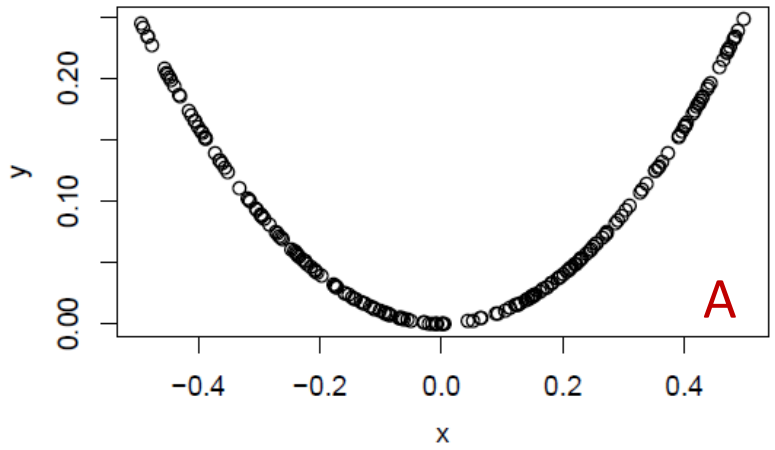
$-1 \leq \rho \leq 1$ where:

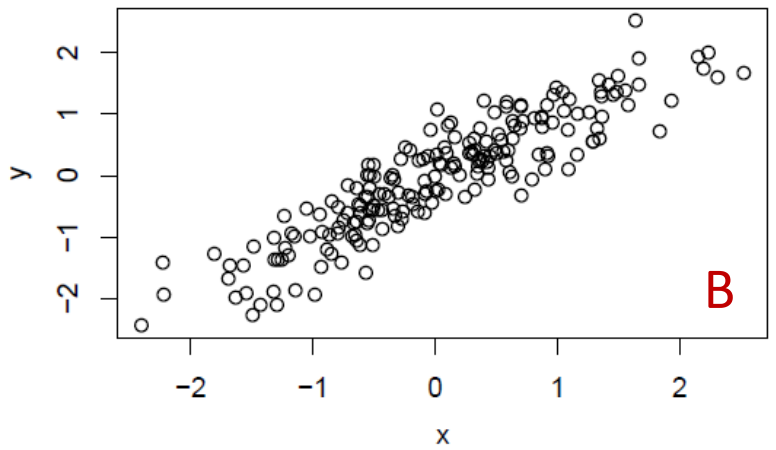
Correlation	Interpretation
ρ > 0	Positive dependency
ρ = 0	No (linear) dependency
ρ < 0	Negative dependency

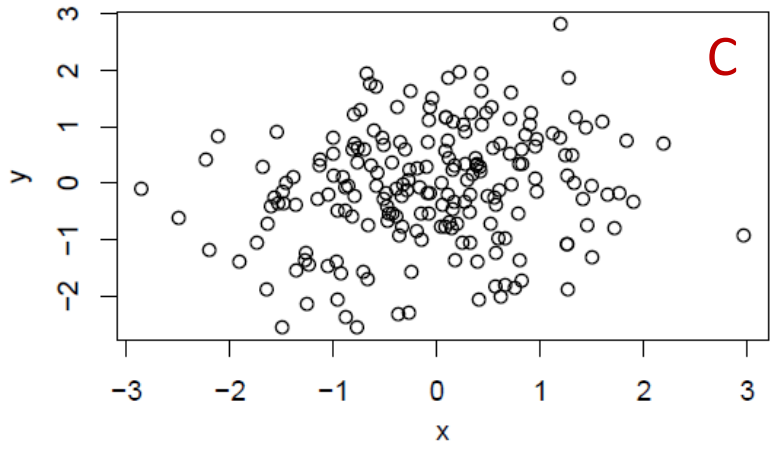


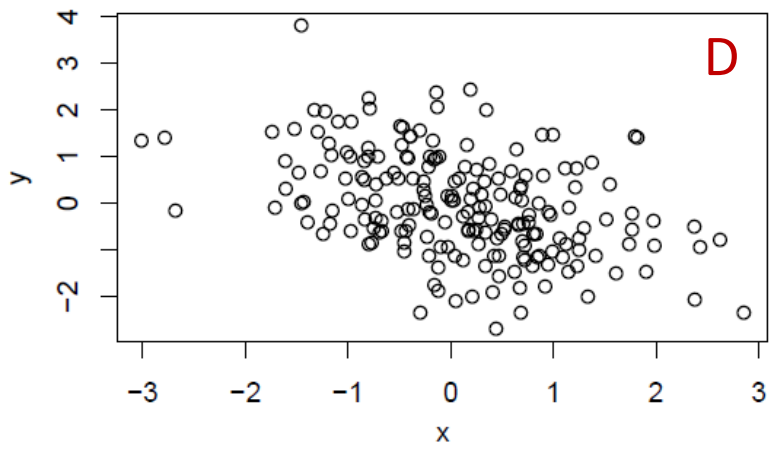
But note: No correlation (i.e. ρ = 0) ⇏ independence (!) (just: no linear dependency).

some examples…





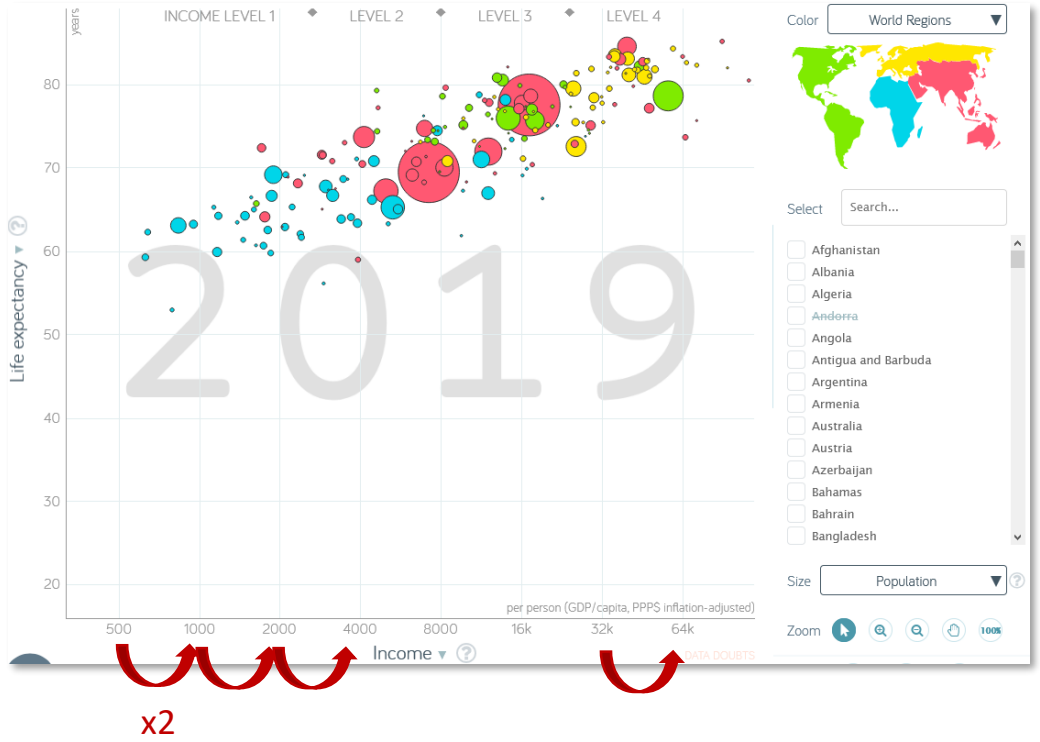




Which is the corresponding plot to a correlation of:

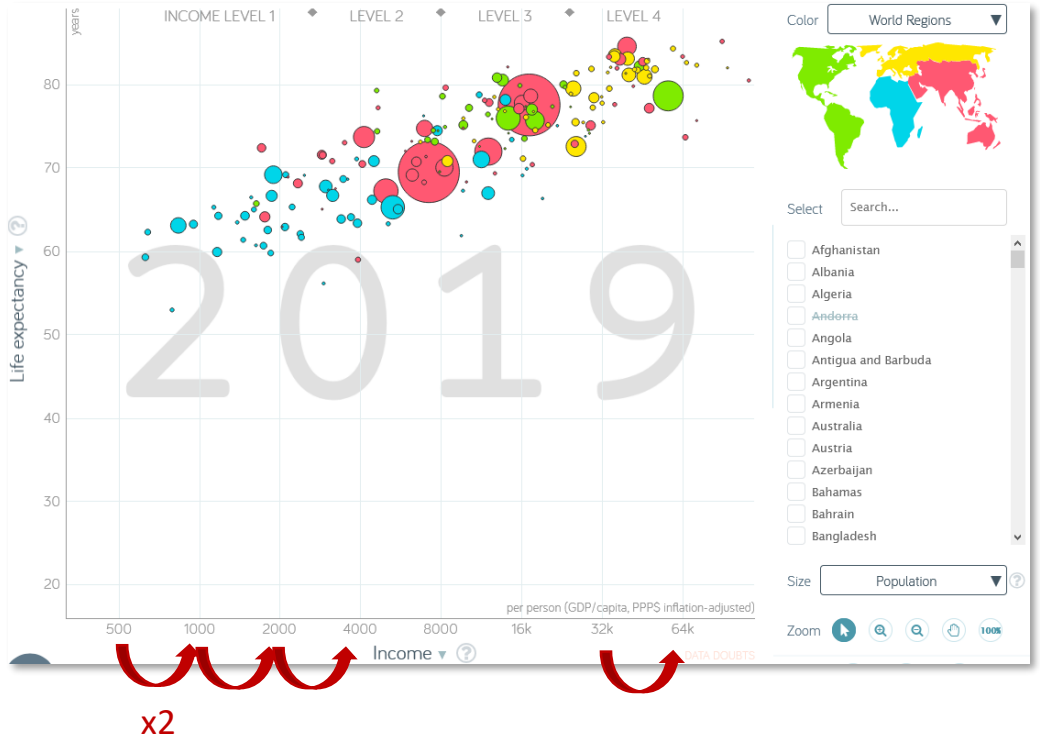
- ❑ **0.9**
- ❑ **-0.42**
- ❑ **0.09**
- ❑ **0.17**

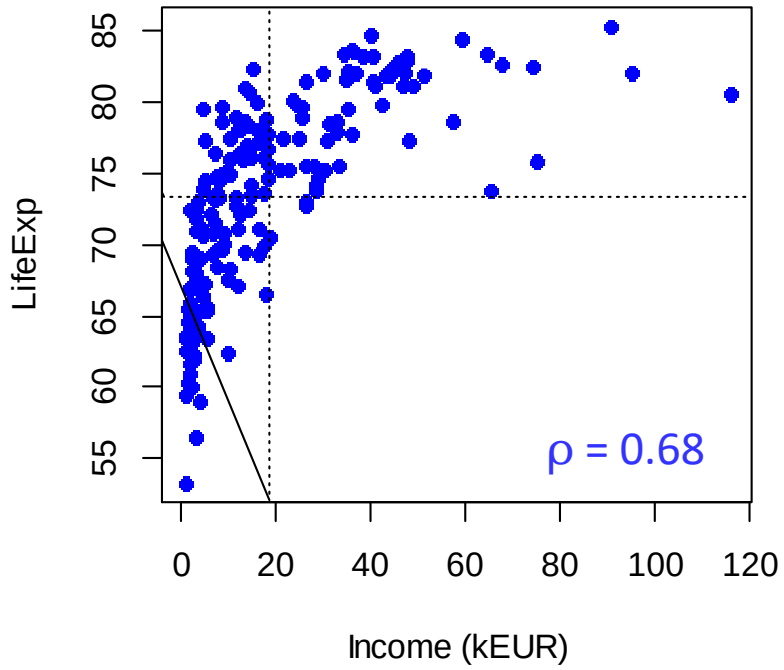
Correlation for Gapminder Data…



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

Correlation for Gapminder Data…

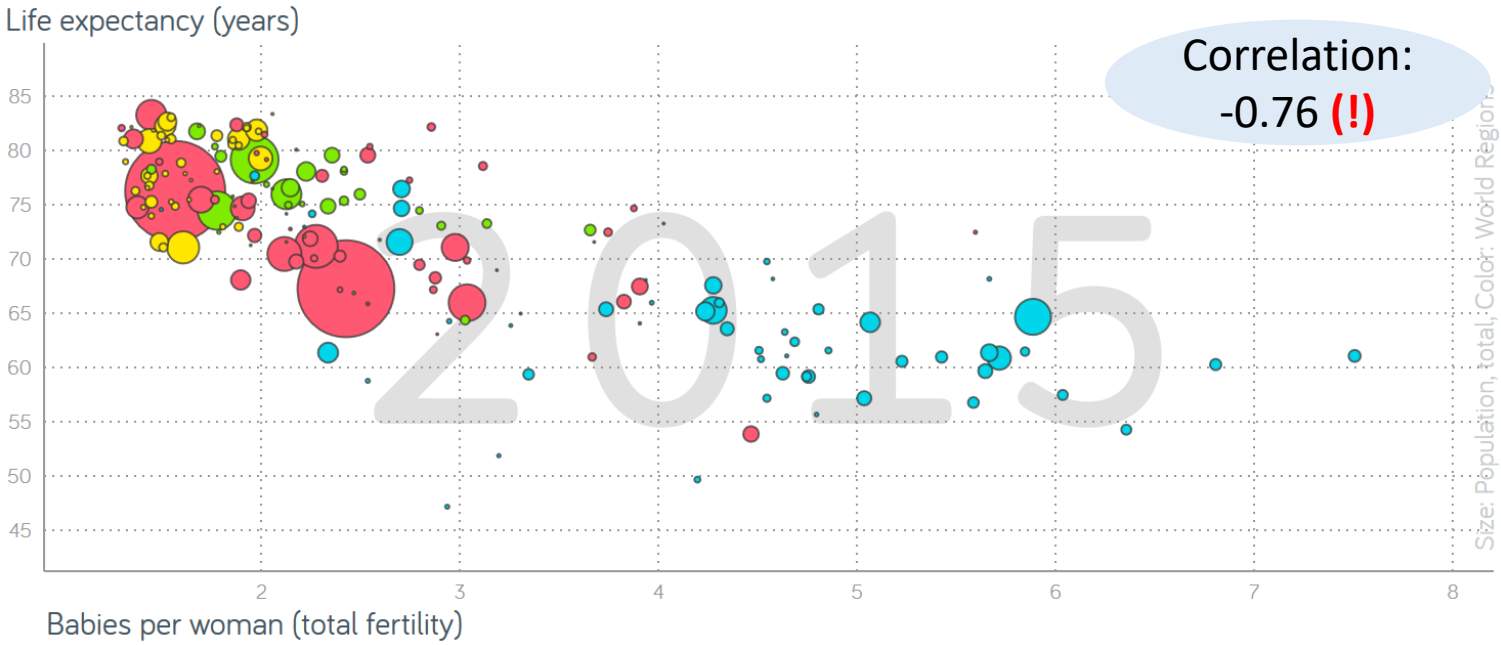




- Note: x-axis logarithmically scaled…
- This is often done if the differences are huge

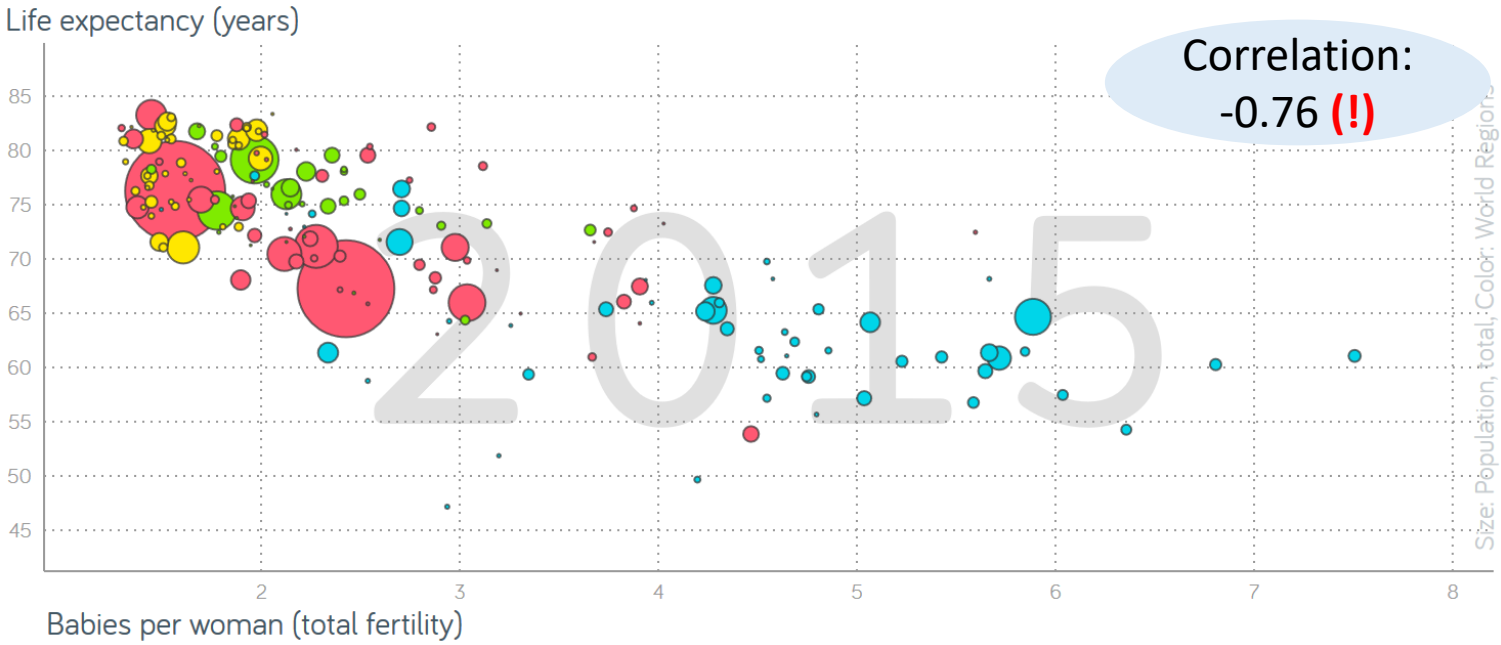
ρ = 0.68 → Strong positive dependency btw income and life expectancy!

A Common Mistake…



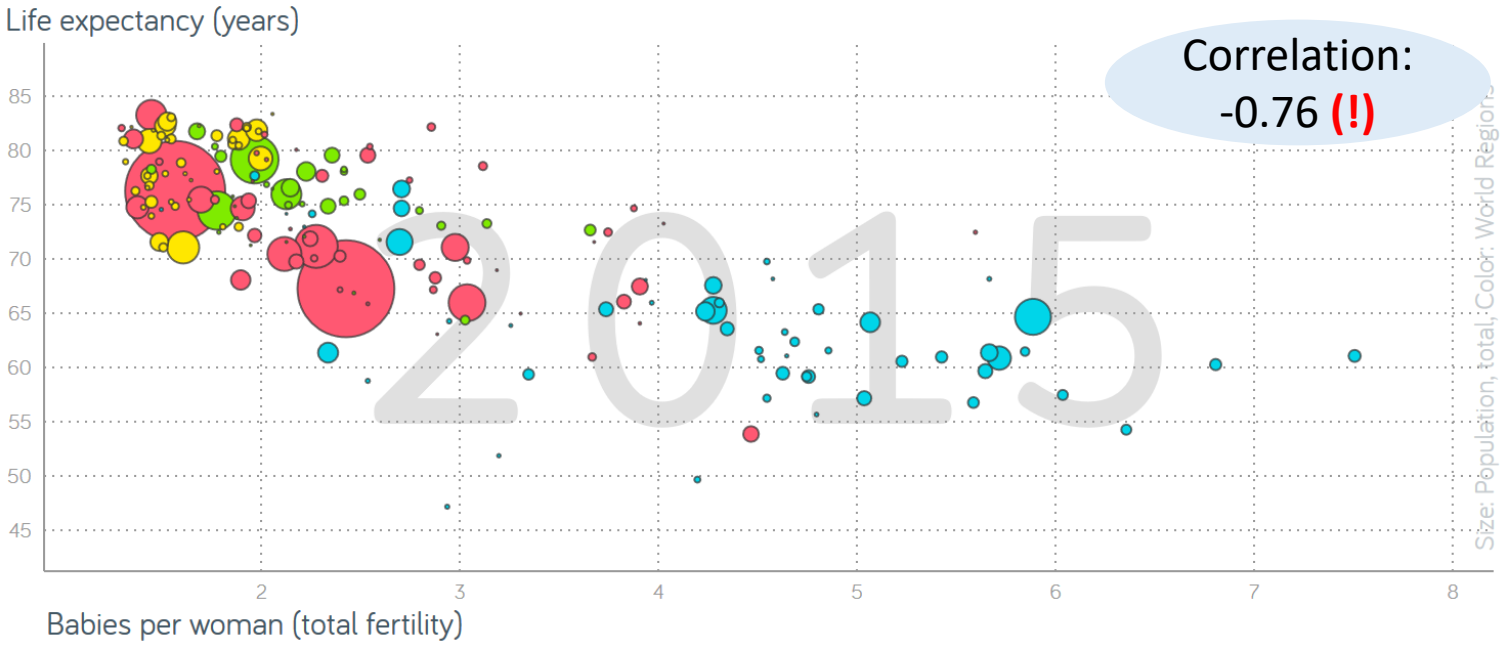
- **What is the interpretation of the negative correlation, here?**

A Common Mistake…



- **What is the interpretation of the negative correlation, here?**
- **Will ‘having children‘ make you die earlier?**

A Common Mistake: Confounder Variables



- **What is the interpretation of the negative correlation, here?**
- **Will 'having children' make you die earlier?**
- **No!** There seems to be a hidden 'confounder' variable behind the data: Richer countries can be assumed to have both, less Ø # of childen and better developed health systems.

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

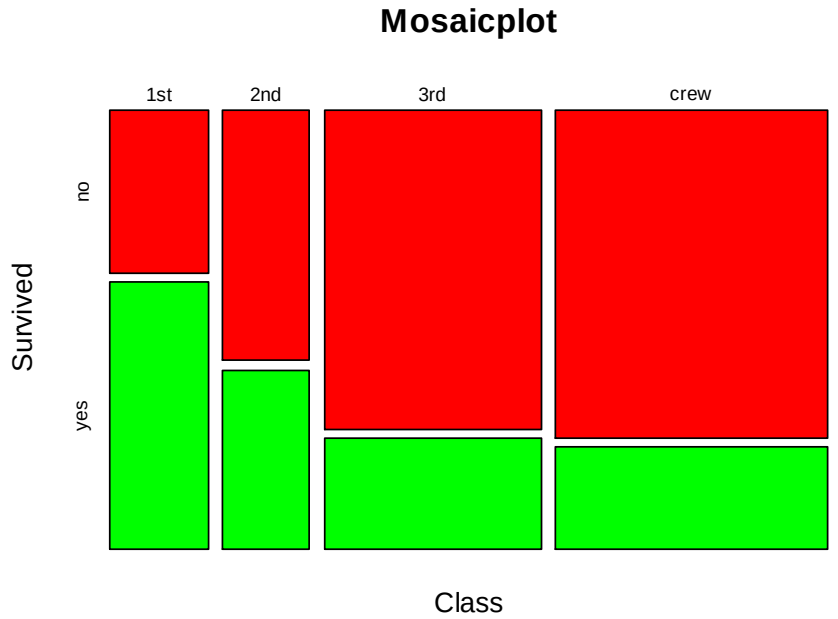


Figure taken from: https://www.geo.de/geolino/mensch/10493-rtkl-geschichte-die-letzte-nacht-auf-der-titanic

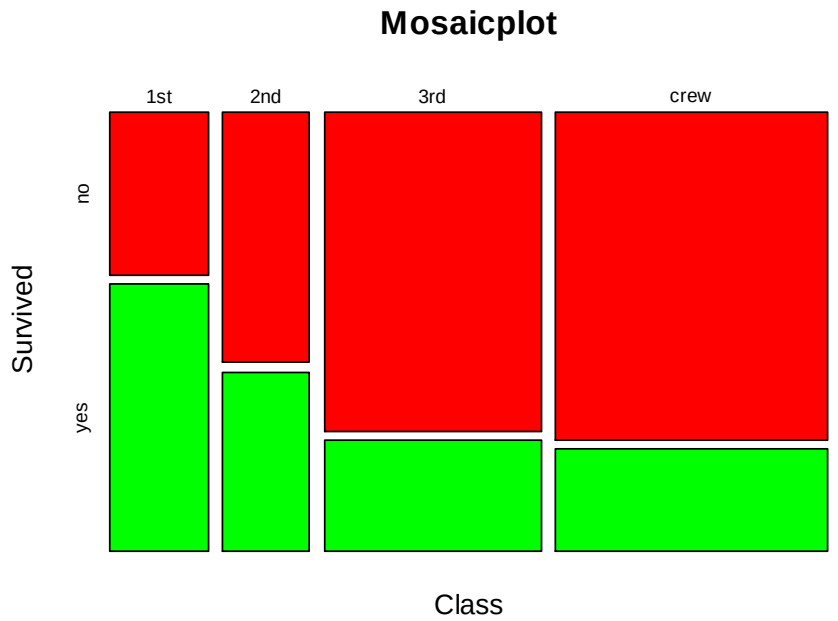
How are the conditional frequencies from the bottom left table are visualized in the mosaic plot?

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			

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

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

Describe what χ^2 measures in a senctence!

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

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

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

0 ≤ V ≤ 1 measures the dependency between two categorical variables.