

Bivariate Statistik

- In der bivariaten Statistik werden **zwei Zufallsvariablen** gemeinsam untersucht, um **Zusammenhänge zu identifizieren**.
 - Gibt es einen Zusammenhang zwischen investierter Lernzeit und Examensnote?
 - Gibt es einen Zusammenhang zwischen der Mathenote im Abitur und der Mathenote an der Hochschule?

Person	Geschlecht	Alter	Mathenote Abi	Mathenote Uni
Ivanka	W	24	2.7	3.3
Leonie	W	18	2.3	2.7
Nico	M	21	1.0	2.3
Janine	W	22	3.0	4.0
Lisa	W	20	3.3	3.0
Mariusz	M	21	2.7	2.7

Rohdaten

Person	Geschlecht	Alter	Mathenote Abi	Mathenote Uni
Ivanka	W	24	2.7	3.3
Leonie	W	18	2.3	2.7
Nico	M	21	1.0	2.3
Janine	W	22	3.0	4.0
Lisa	W	20	3.3	3.0
Mariusz	M	21	2.7	2.7

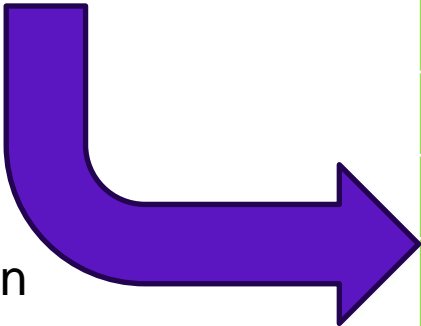
```
import pandas as pd

pd.crosstab()
```

Kreuz- / Kontingenztafel für Alter / Geschlecht

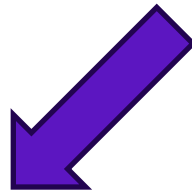
Alter	18-19	20-21	22-23	24-25	gesamt
Geschlecht					
W	1	1	1	1	4
M	0	2	0	0	2
gesamt	1	3	1	1	6

Häufigkeit der Elemente zählen



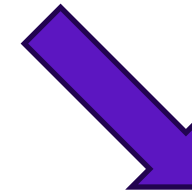
Kontingenztafel Alter / Geschlecht

Alter	18-19	20-21	22-23	24-25	gesamt
Geschlecht					
W	1	1	1	1	4
M	0	2	0	0	2
gesamt	1	3	1	1	6



Spaltenprozent

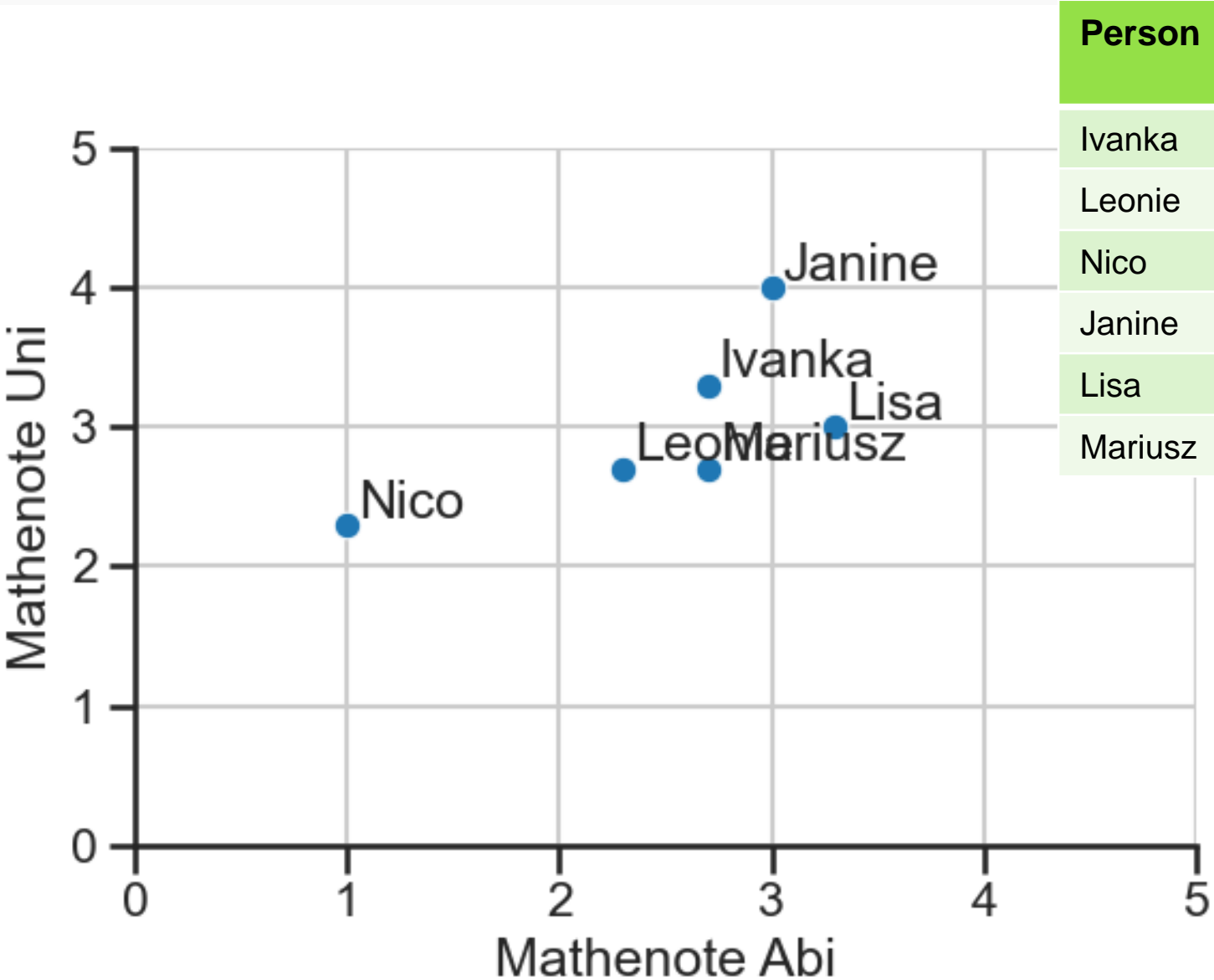
Alter	18-19	20-21	22-23	24-25
Geschlecht				
W	100%	33%	100%	100%
M	0	67%	0	0
gesamt	100%	100%	100%	100%



Zeilenprozent

Alter	18-19	20-21	22-23	24-25	gesamt
Geschlecht					
W	25%	25%	25%	25%	100%
M	0	100%	0	0	100%

- Streuungsdiagramme dienen der einfachen Veranschaulichung des Zusammenhangs zweier Variablen.
- Bei Streuungsdiagrammen müssen x und y idR mindestens Ordinalskala haben. Nominalskalen sind nicht immer möglich. Meistens werden bei Scatterplots zumindest Intervallskalierte Daten verwendet.
- Verschiedene Markersymbole oder Markerfarben können genutzt werden, um bis zu zwei weitere Variablen grafisch darzustellen.



Person	Geschlecht	Alter	Mathenote Abi	Mathenote Uni
Ivanka	W	24	2.7	3.3
Leonie	W	18	2.3	2.7
Nico	M	21	1.0	2.3
Janine	W	22	3.0	4.0
Lisa	W	20	3.3	3.0
Mariusz	M	21	2.7	2.7

7

Zusammenhangs- maße

Varianz - (Quadratische) Streuung einer einzigen Variablen X

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

NEU: Kovarianz – gemeinsame Streuung zweier Variablen X, Y

$$\text{cov}_{XY} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Mindestens Intervall- oder Verhältnisskala sind erforderlich.

Teile die Summe durch den Stichprobenumfang $n - 1$

Berechne die Abweichung der Einzelbeobachtungen x_i und y_i von ihren arithmetischen Mitteln \bar{x} und \bar{y}

$$cov_{XY} = \frac{1}{n - 1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Kovarianz (Symbol cov oder c) der Variablen X und Y

Zähle von 1 bis n , merke die Zahl als i , und summiere für jeden Index i das Ergebnis der Rechnung rechts

- Ist $c_{XY} > 0$ so existiert ein positiver linearer Zusammenhang zwischen den Variablen
- Ist $c_{XY} < 0$ so existiert ein negativer linearer Zusammenhang zwischen den Variablen

ACHTUNG: Die Skala der Kovarianz ist nicht standardisiert und schwankt stark abhängig von den gemessenen Variablen. Je größer die Ausgangswerte desto größer die Kovarianz. Die Stärke der Kovarianz ist daher nicht anhand ihres Wertes interpretierbar.

LÖSUNG: Eine standardisierte Form der Kovarianz, die Korrelation.

```
Python:  
import statistics  
statistics.covariance(x,y)  
  
import pandas as pd  
df.cov()
```

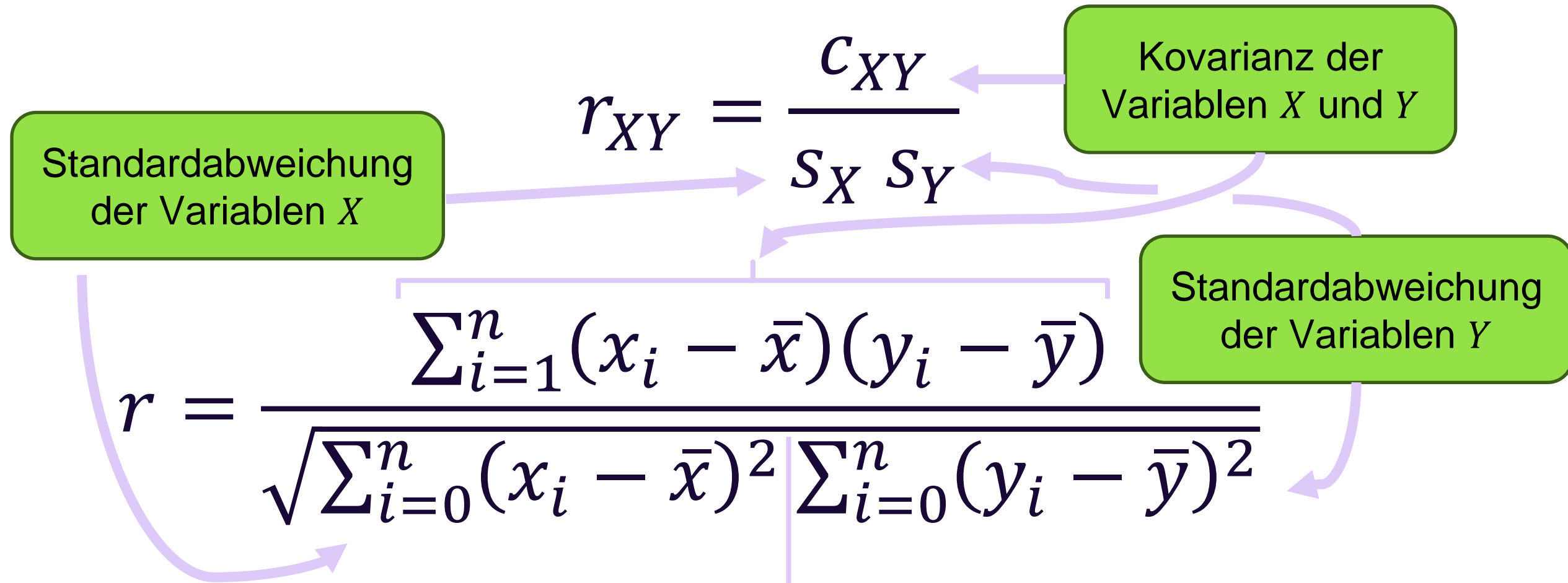
Korrelation = Maßzahl für den linearen Zusammenhang zweier Variablen

- Die Werte liegen zwischen -1 und 1
 - 1 bedeutet perfekter positiver linearer Zusammenhang, d.h. die Punkte liegen auf einer aufsteigenden Geraden
 - -1 bedeutet perfekter negativer linearer Zusammenhang, d.h. die Punkte liegen auf einer absteigenden Geraden
 - 0 bedeutet keinen linearen Zusammenhang
- Pearsons Korrelationskoeffizient (Pearsons r) darf für metrische Variablen berechnet werden

Python:

```
import statistics  
statistics.correlation(x,y)
```

```
import pandas as pd  
df.corr()
```



Standardabweichung der Variablen X

Kovarianz der Variablen X und Y

Standardabweichung der Variablen Y

$$r_{XY} = \frac{c_{XY}}{s_X s_Y}$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=0}^n (x_i - \bar{x})^2 \sum_{i=0}^n (y_i - \bar{y})^2}}$$

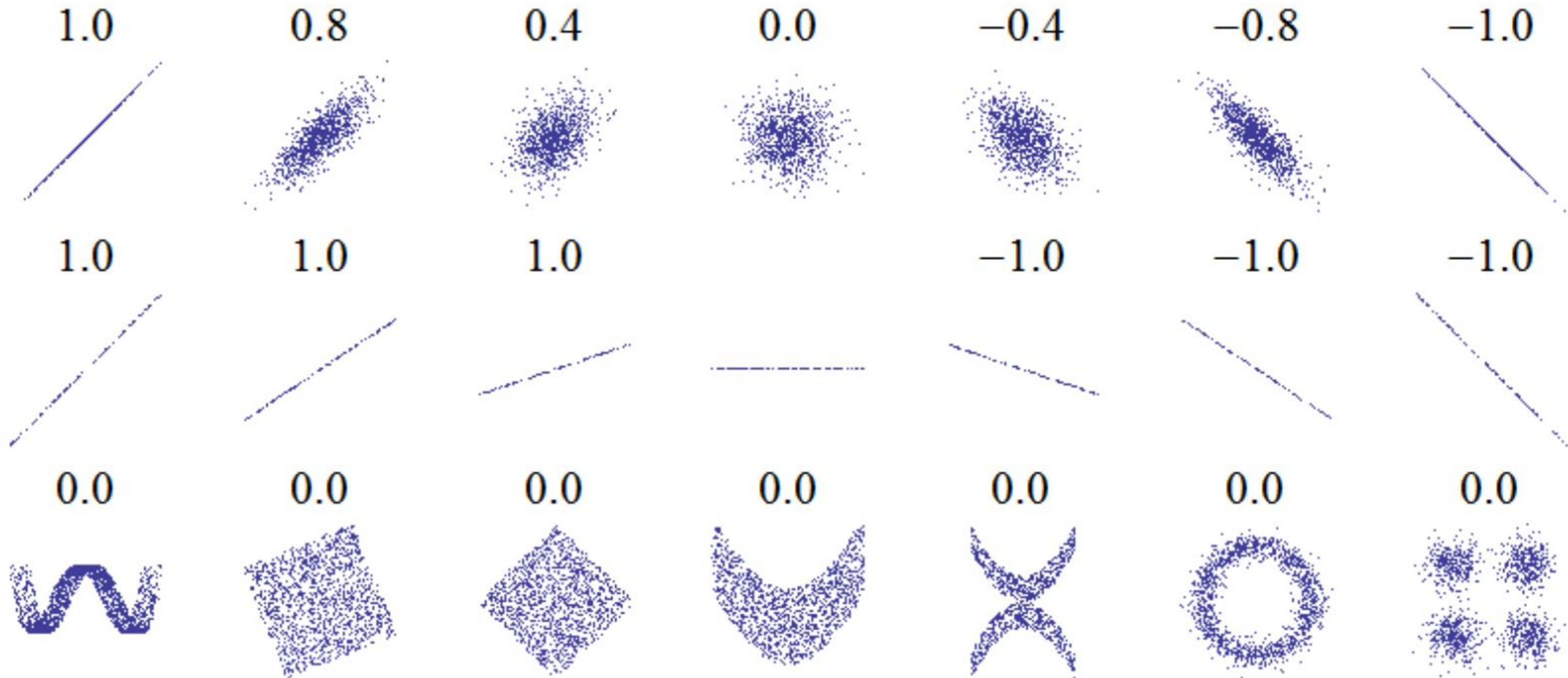
Die Standardisierung der Skala der Korrelation ermöglicht es, Korrelationsstärken zu vergleichen. Eine Korrelation r von 0.86 ist statistisch stärker als eine Korrelation von 0.45.

Korrelation kann auch in ihrer Stärke kategorisiert werden, z.b. mit:

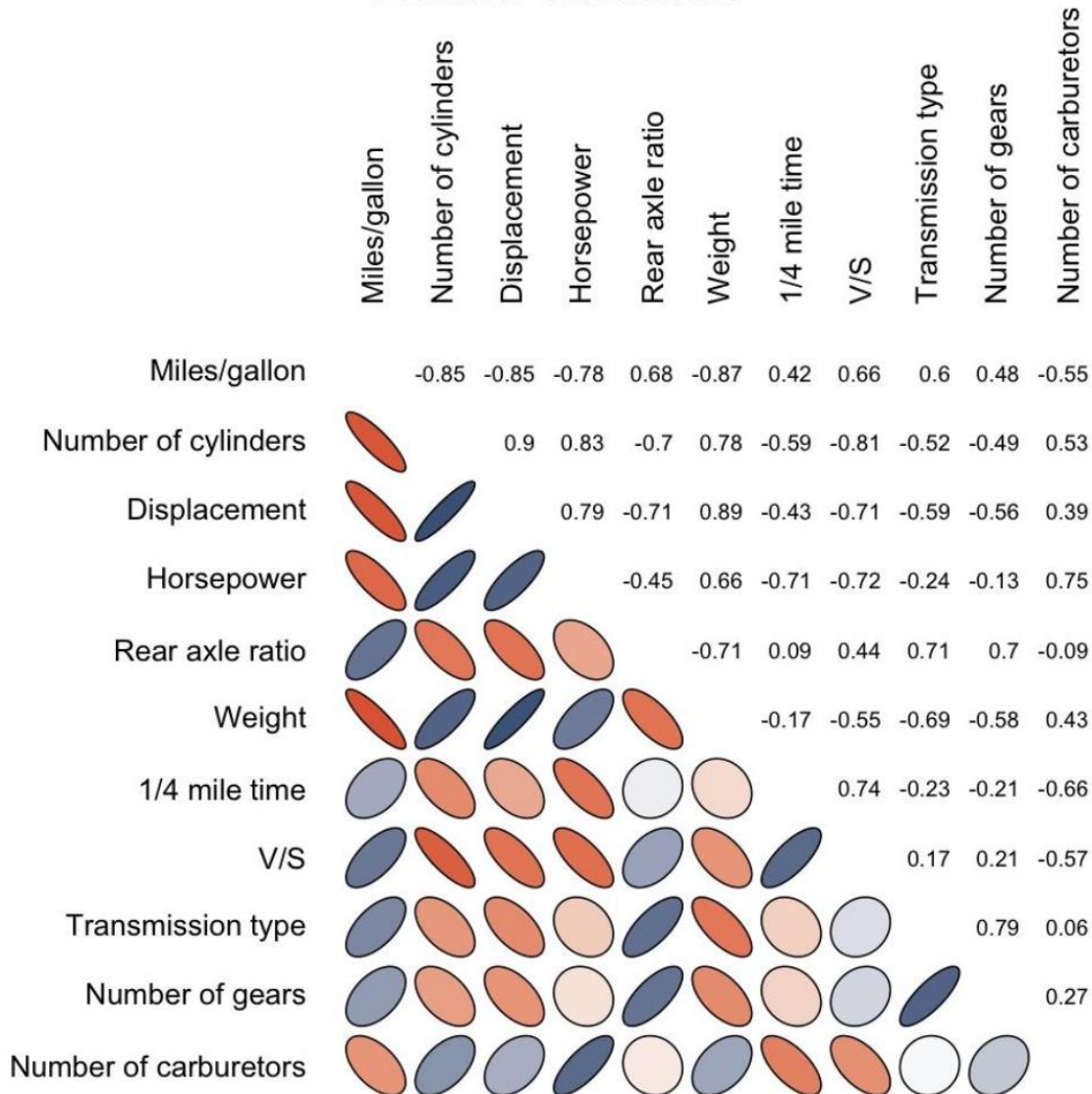
$$0 < |r_{XY}| < 0.\bar{3} = \text{“}schwach\text{“}$$

$$0.\bar{3} < |r_{XY}| < 0.\bar{6} = \text{“}mittel\text{“}$$

$$0.\bar{6} < |r_{XY}| < 1 = \text{“}stark\text{“}$$



Predictor correlations



Korrelationsmatrix

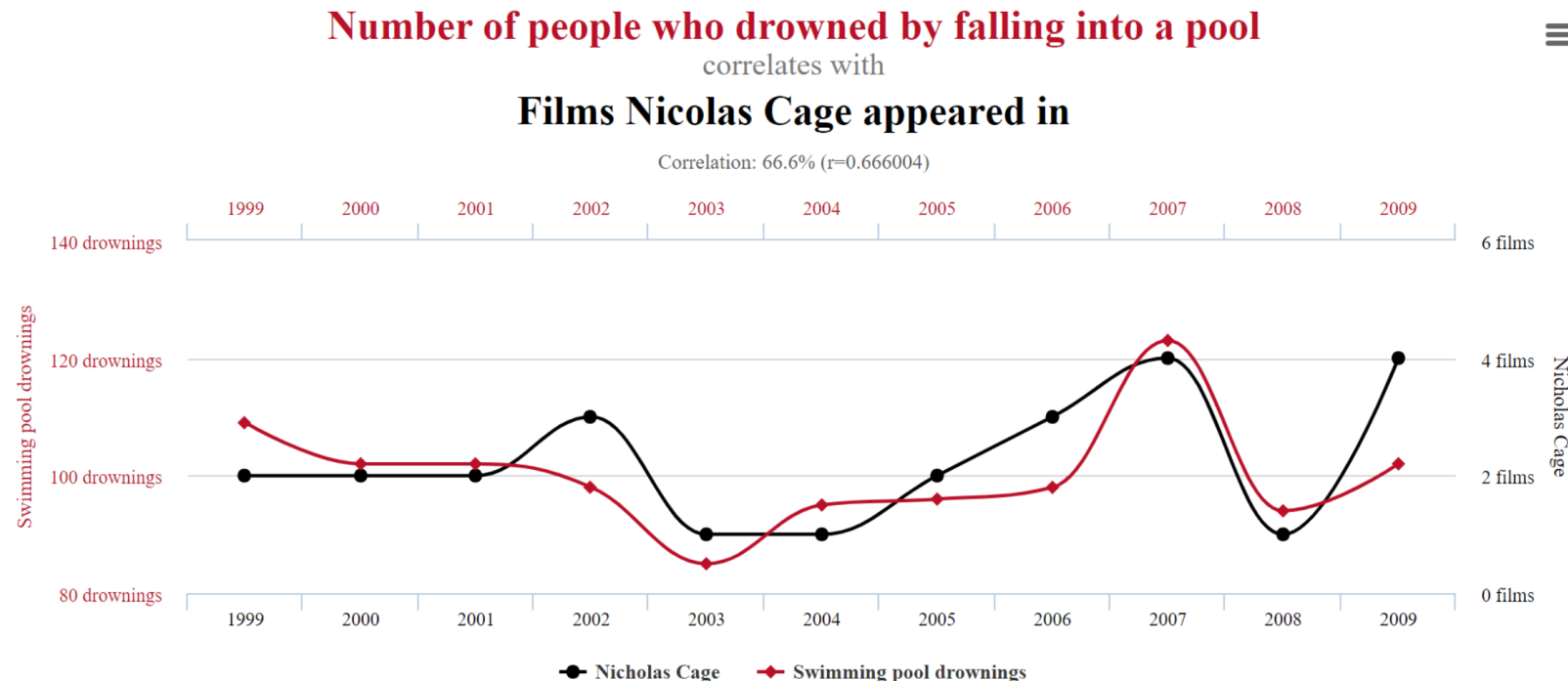
- Korrelationen aller numerischer Variablen eines Datensatzes (hier mpg) gegeneinander.
- Farbe signalisiert Stärke und Richtung der Korrelation (divergente Farbskala)
- Form repräsentiert Form der Punktwolke

<https://hlplab.files.wordpress.com/2012/03/my-plotcorr.jpg>

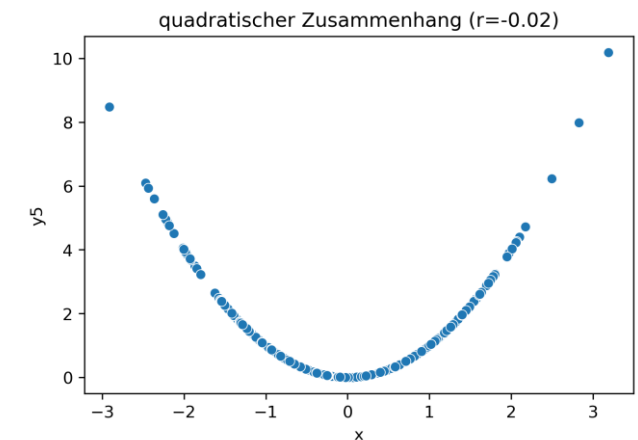
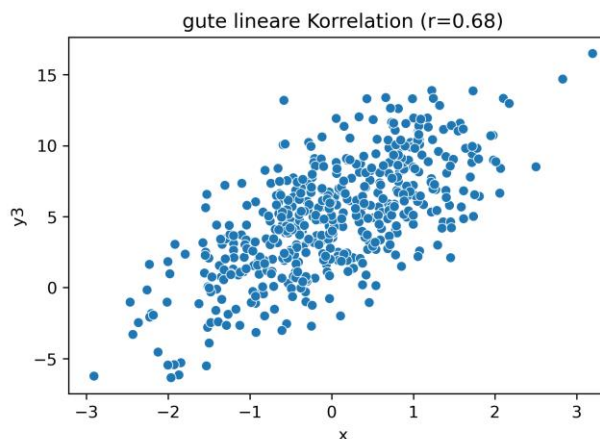
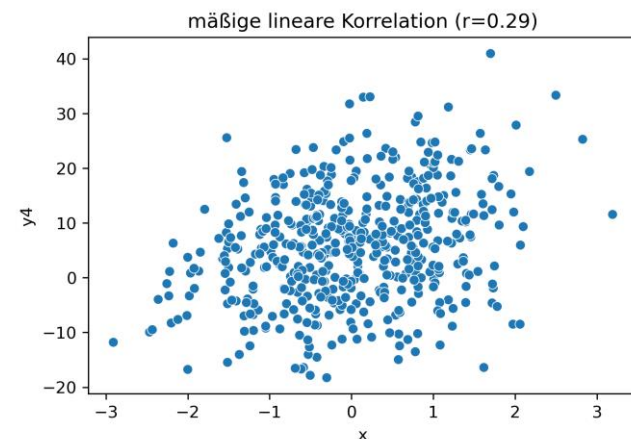
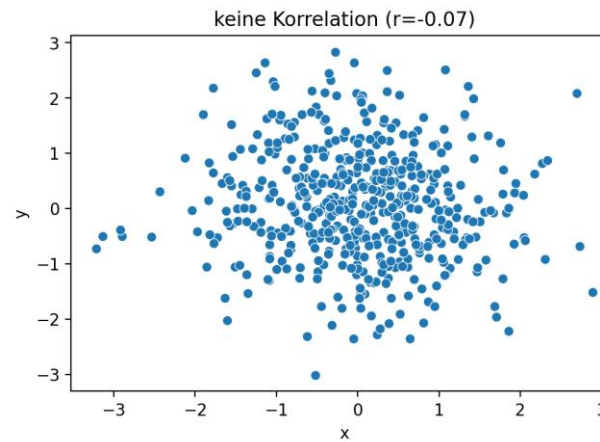
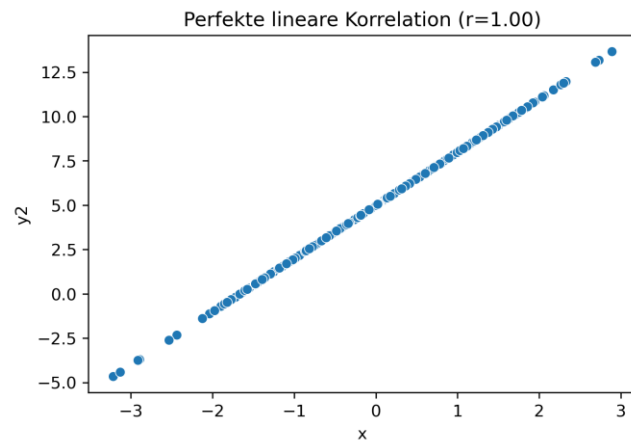
Achtung: Nur weil statistisch ein starker Zusammenhang besteht, muss nicht tatsächlich ein sachlicher Zusammenhang vorliegen. Branchenwissen kann weiterhelfen, Sachverhalte einzuschätzen.

Kinder, die zu Hause frühstücken, zeigen bessere Leistungen in der Schule

Menschen, die Ihre Schuhe beim Schlafen anbehalten, haben ein erhöhtes Risiko, mit Kopfschmerzen aufzuwachen



Formel:
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$



Nur lineare Korrelationen
werden richtig erfasst.

Für ordinale Variablen gibt es auch Möglichkeiten. Diese nennt man **Rangkorrelationskoeffizienten**.

Diese vergleichen den Rang der beobachteten Werte, also die Position in einer geordneten Liste. Dabei geht es nicht mehr um den linearen Zusammenhang, sondern ob die Ränge beider Variablen aufsteigen.

Die zwei bekanntesten Rangkorrelationskoeffizienten sind

- Spearmans ρ (rho)
- Kendalls τ (tau)

Python:

```
import pandas as pd
df.corr(method="spearman")
df.corr(method="kendall")
```