

Statistik für Data Scientists

Vorlesung 4: Korrelation & Zusammenhang

Prof. Dr. Siegfried Handschuh
DS-NLP
Universität St. Gallen

Recap & Ziele heute

- Letzte Woche : **Deskriptive Statistik & Visualisierung.**
- Heute: **Zusammenhangsmasse** (Pearson, Spearman, Kendall).
- **Simpson-Paradox & Konfundierung – Achtung!**
- Visualisierung von Korrelationen.
- Praxisfallen & Interpretation

Einstieg & Intuition

Warum wir Korrelation brauchen

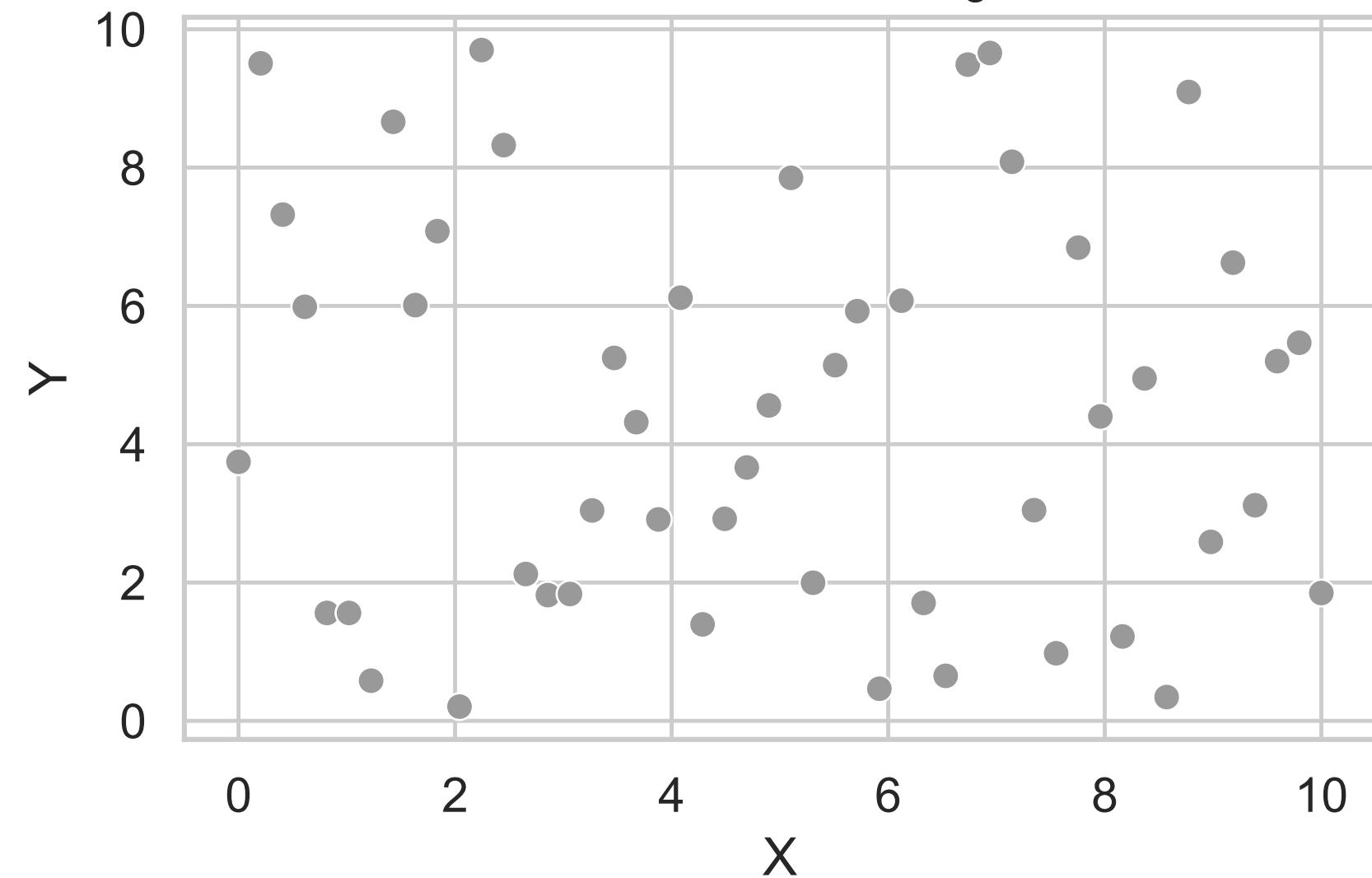
Ändert sich Y , wenn sich X ändert?

- Korrelation = deine **erste quantitative Antwort** auf diesen Zusammenhang.
- Beispiel: Temperatur $\uparrow \Rightarrow$ Eisverkauf \uparrow . Ein Muster, das messbar ist.
- Aber: Das ist **noch keine Kausalität**, nur ein messbares Muster.

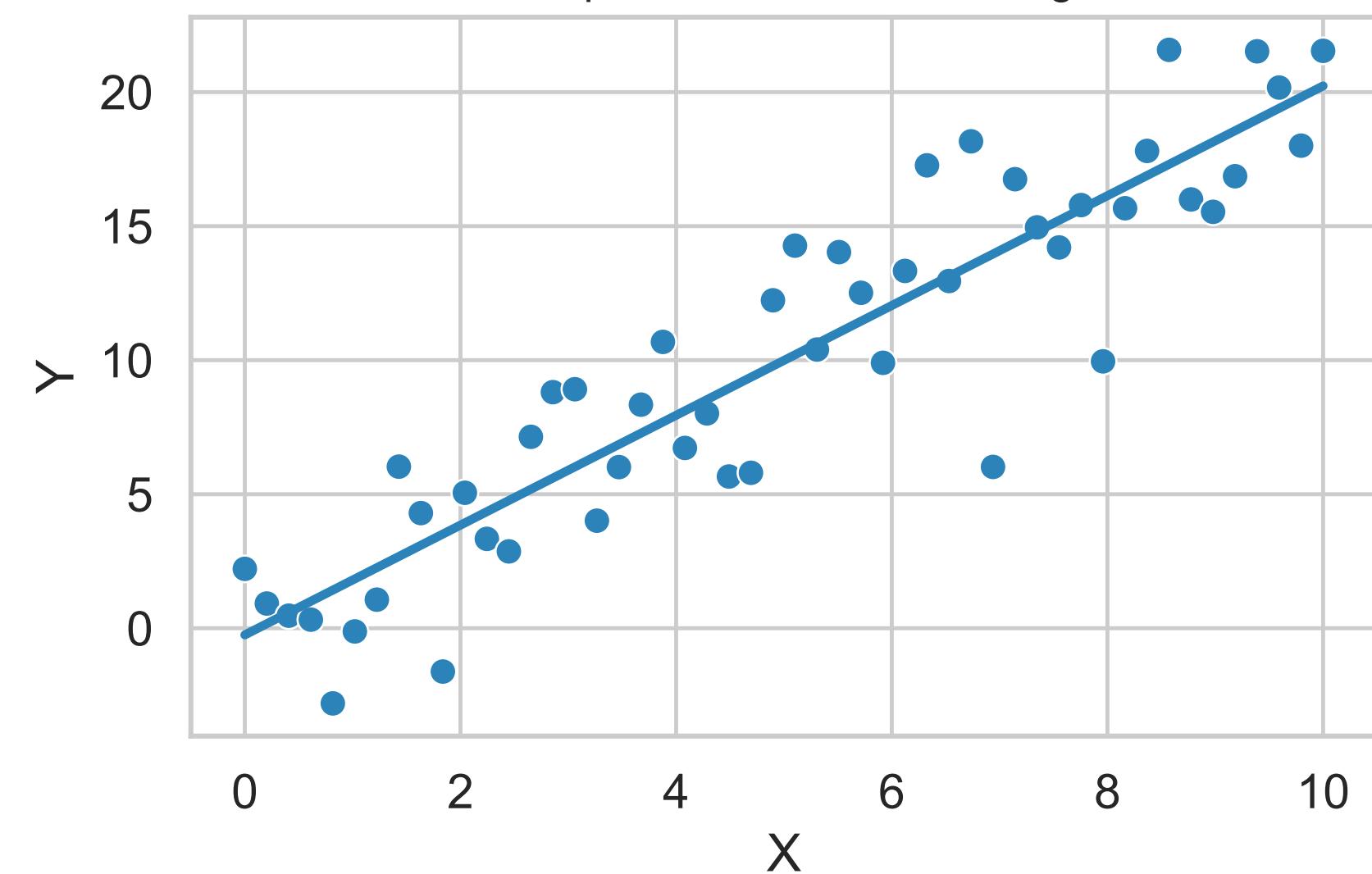
Mini-Check: Warum reicht es nicht, nur Mittelwerte zu vergleichen?

Warum wir Korrelation brauchen

Kein Zusammenhang



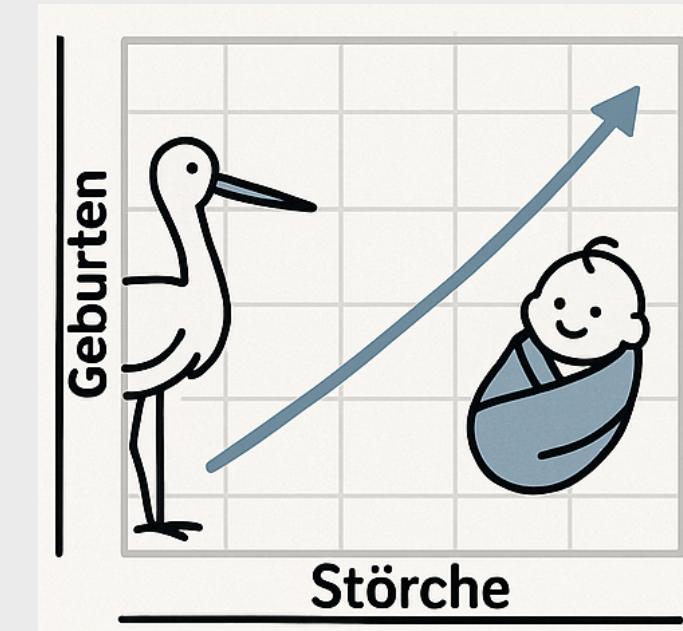
Klar positiver Zusammenhang



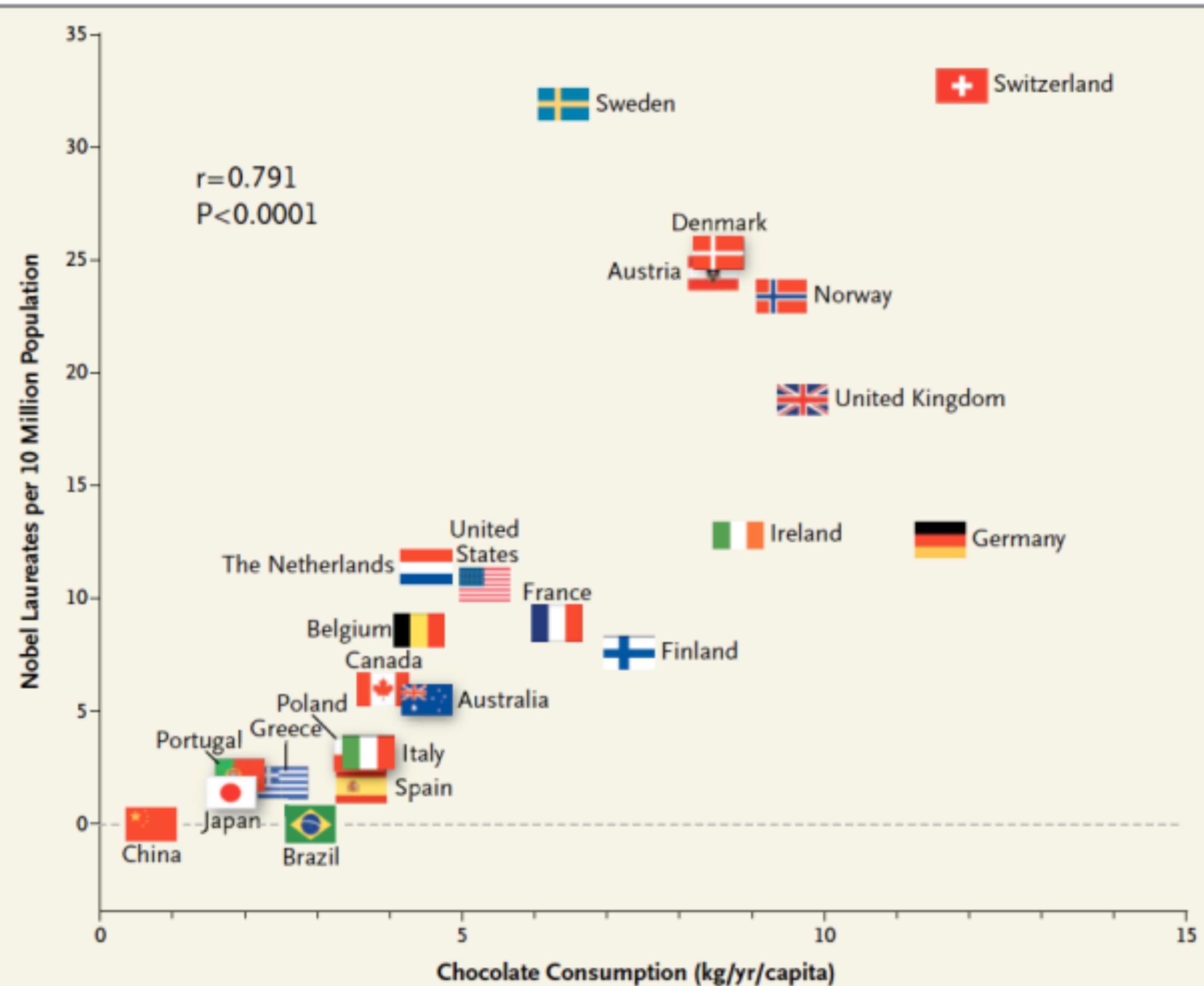
Alltagsintuition

- Du erkennst Muster intuitiv. Statistik macht sie **überprüfbar** und **verifizierbar**.
- Denk an: «Mehr Training ⇒ bessere Leistung?» Das fühlt sich richtig an.
- Deine Wahrnehmung ist visuell, die Statistik macht es **quantitativ**.

Mini-Check: Nenne ein Beispiel für eine scheinbare, aber trügerische Korrelation.



- Der Unterschied zwischen «**scheint so**» und «**ist so**».
- Wie stark ist dieser Trend und in welche Richtung geht er?



Was heisst «Zusammenhang»?

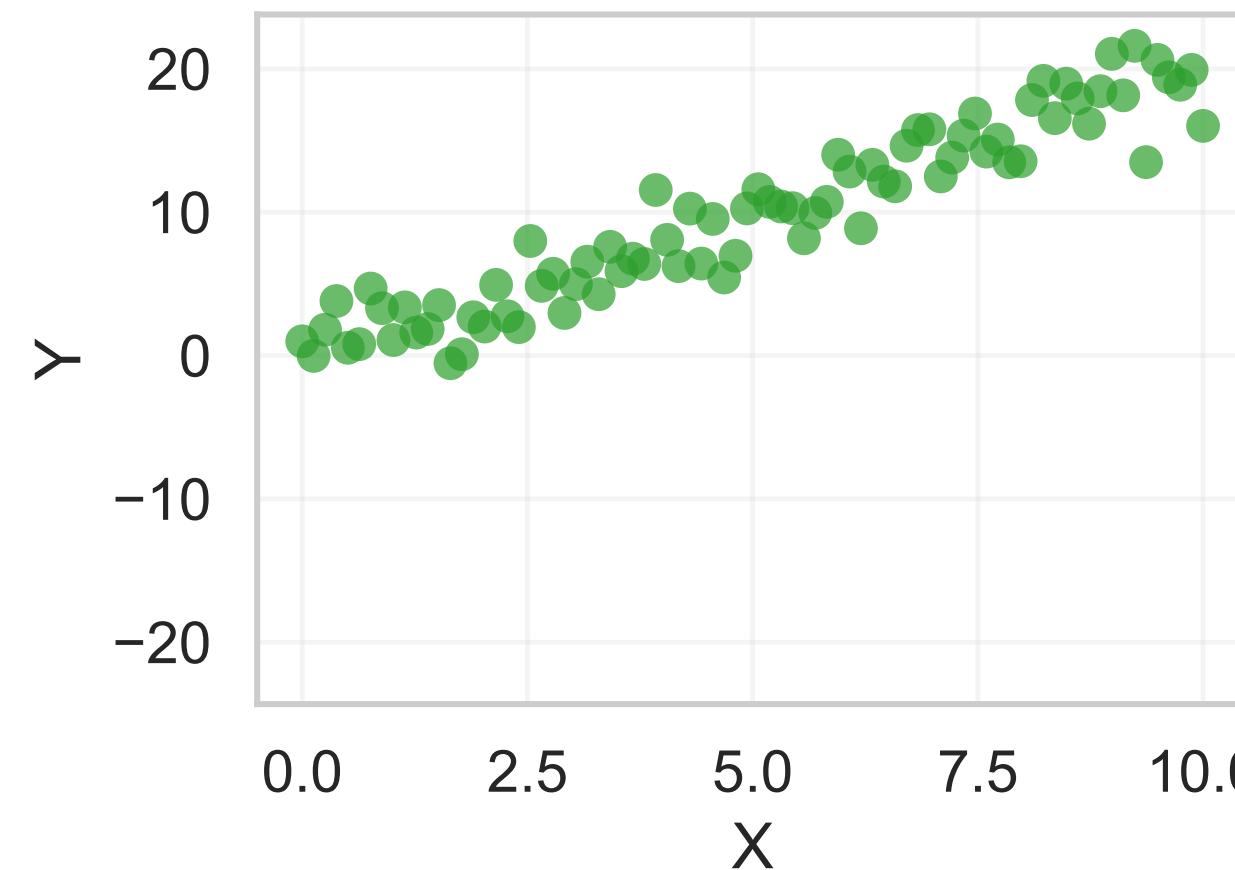
Definition: Zwei Variablen «hängen zusammen», wenn Veränderungen in X mit systematischen Veränderungen in Y einhergehen.

- Richtungen: **positiv**, **negativ**, **kein Zusammenhang**.
- Achtung: Es gibt **Linear** vs. **nichtlinear**. Das ist entscheidend für die Wahl deines Messinstrumentes.
- Formales Denken \Rightarrow Kovarianz, Korrelation.

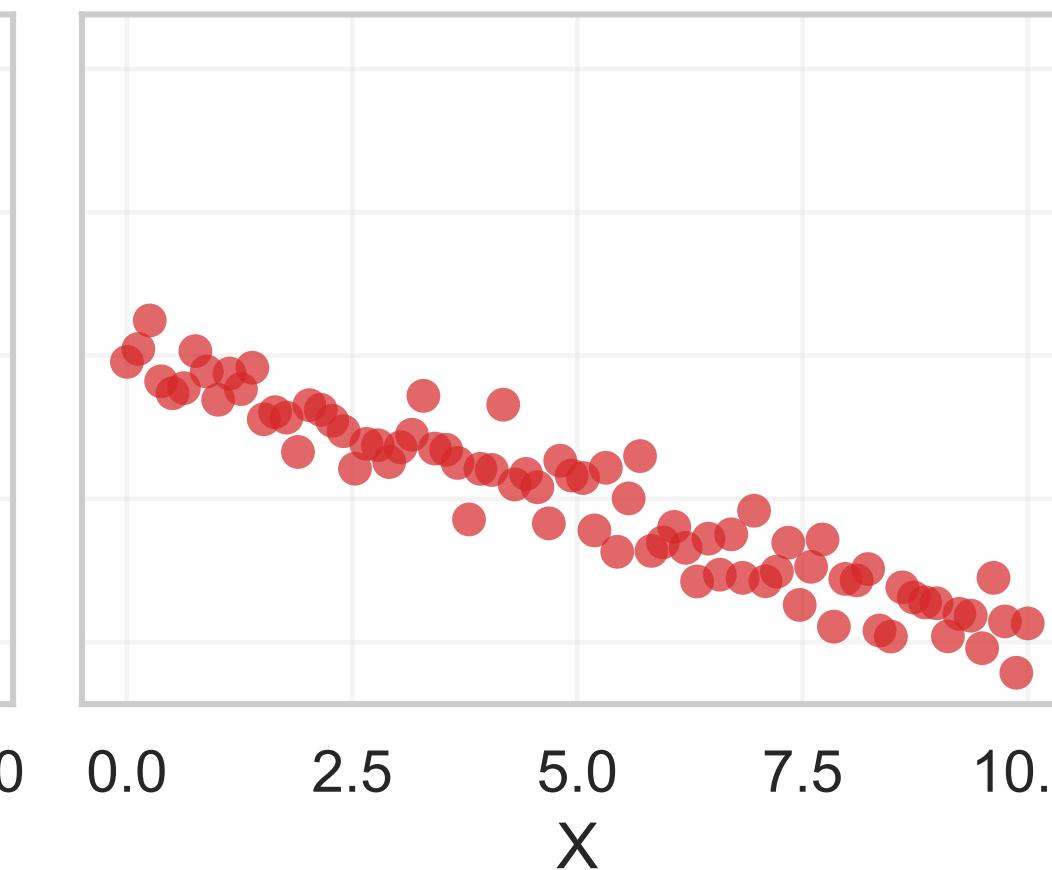
Mini-Check: Was unterscheidet «positiv» von «negativ» im Streudiagramm?

Richtung des Zusammenhangs

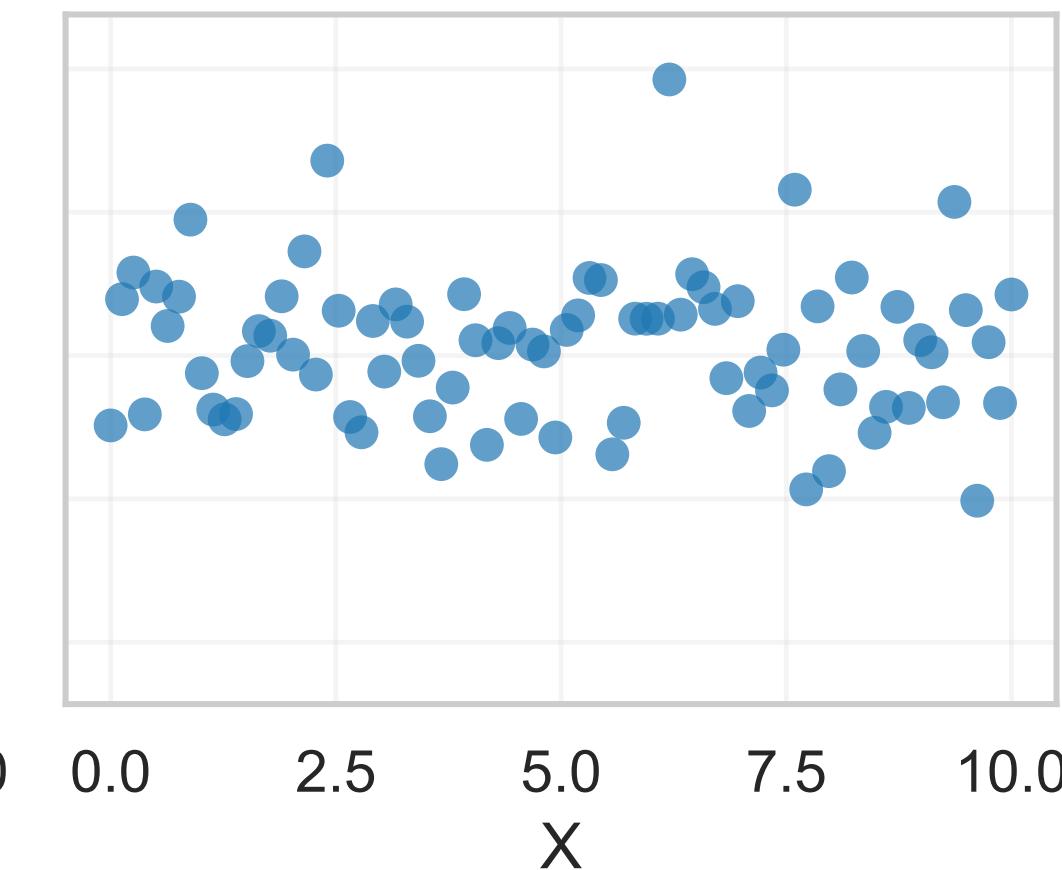
Positiver Zusammenhang



Negativer Zusammenhang



Kein Zusammenhang



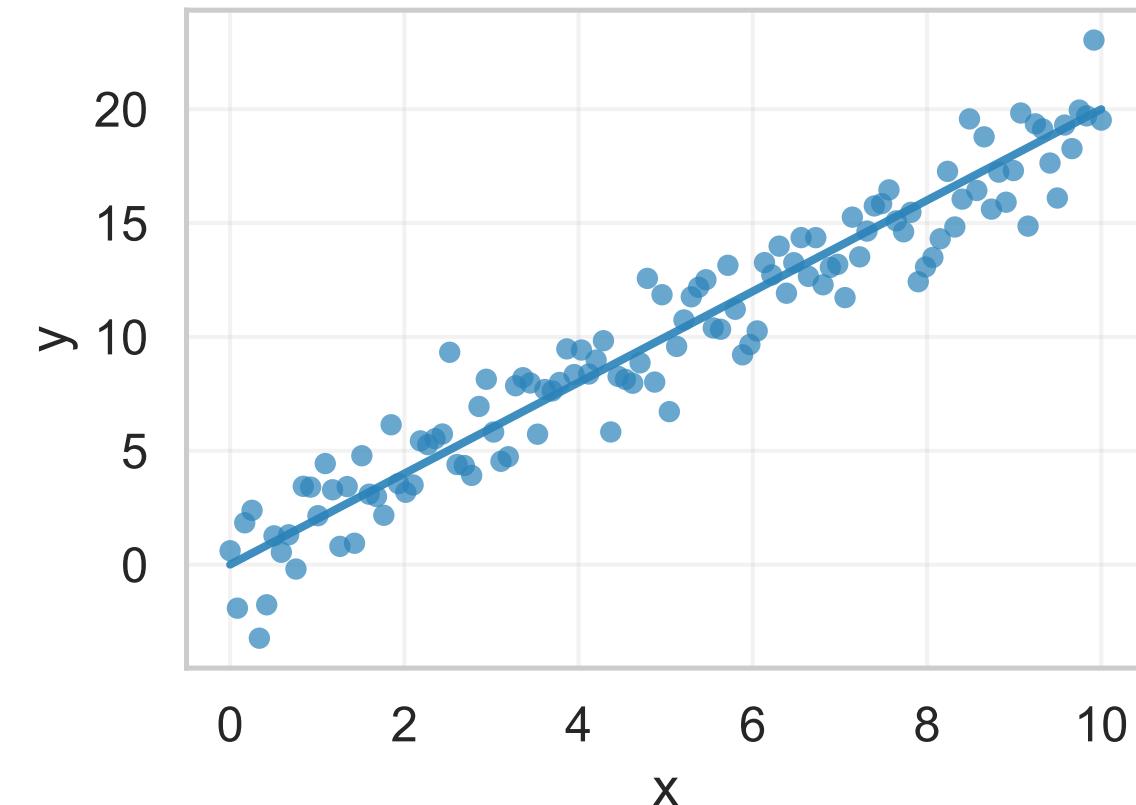
Linear vs. Monoton

- **Linear:** $y \approx a + bx$. (Die Steigung ist überall gleich, wie eine Rampe.)
- **Monoton:** $x \uparrow \Rightarrow y$ immer \uparrow (aber die Steigung kann sich ändern: es ist eine Kurve, die nur nach oben geht).
- Beispiel: Stress $\uparrow \Rightarrow$ Leistung \uparrow bis Plateau \Rightarrow Abfall. (DAS ist **nicht** monoton).
- Das ist **entscheidend** für die Wahl deines Korrelationsmaßes (Pearson vs. Spearman).

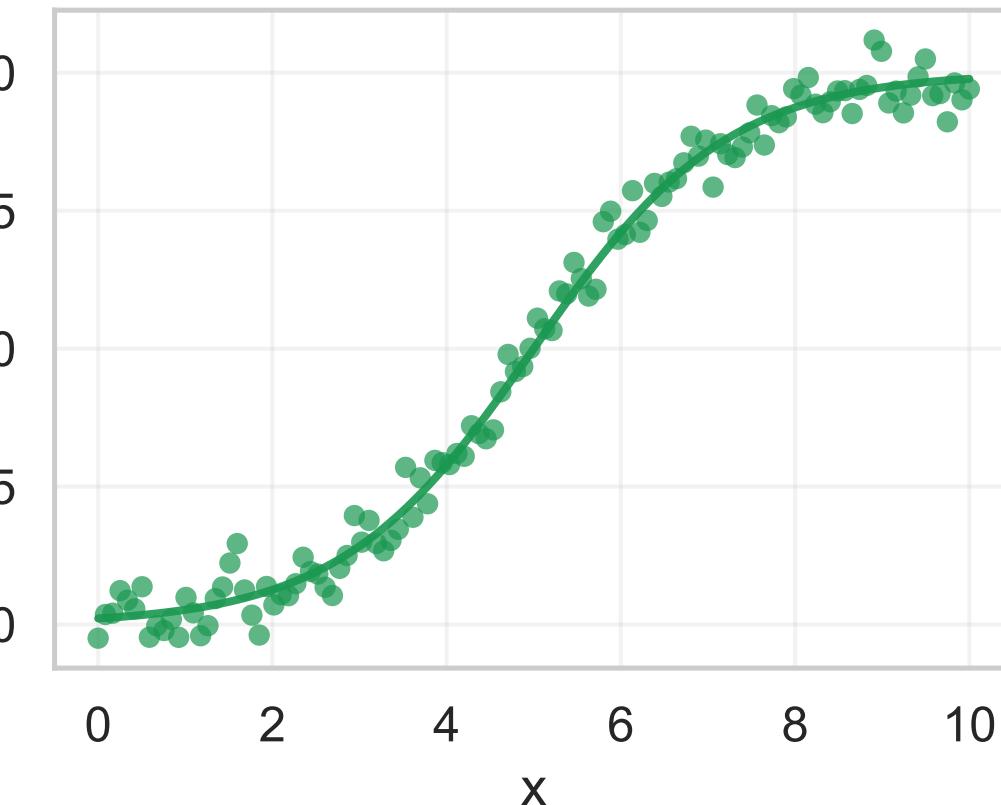
Mini-Check: Zeichne auf Papier ein monoton nicht-lineares Beispiel.

Linear vs. Monoton – Form des Zusammenhangs ist entscheidend

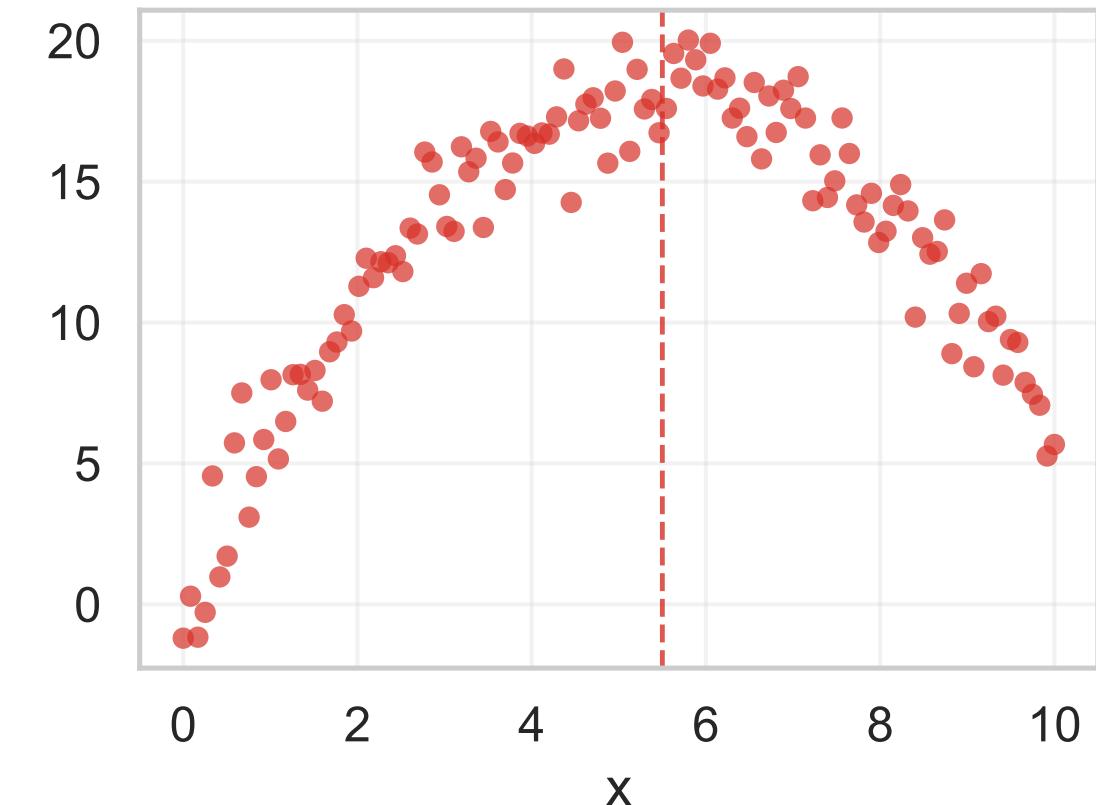
Linear: $y \approx a + bx$



Monoton, nicht linear (steigt, flacht ab)



Nicht monoton (Plateau/Abfall)



Korrelation: Richtung und Stärke

Koeffizient r lebt in $[-1, 1]$. $r > 0 \Rightarrow$ positiv, $r < 0 \Rightarrow$ negativ.

- $|r| \approx 0 \Rightarrow$ kein linearer Zusammenhang.
- Beispiele: $r = 0.9$ (stark positiv); $r = -0.5$ (mittel negativ).
- **Achtung:** Der r -Wert ist nicht linear interpretierbar! $r = 0.7$ heisst nicht «70 % Erklärung».

Mini-Check: Was heisst $r = 0$?

Unterschied zwischen r und r^2

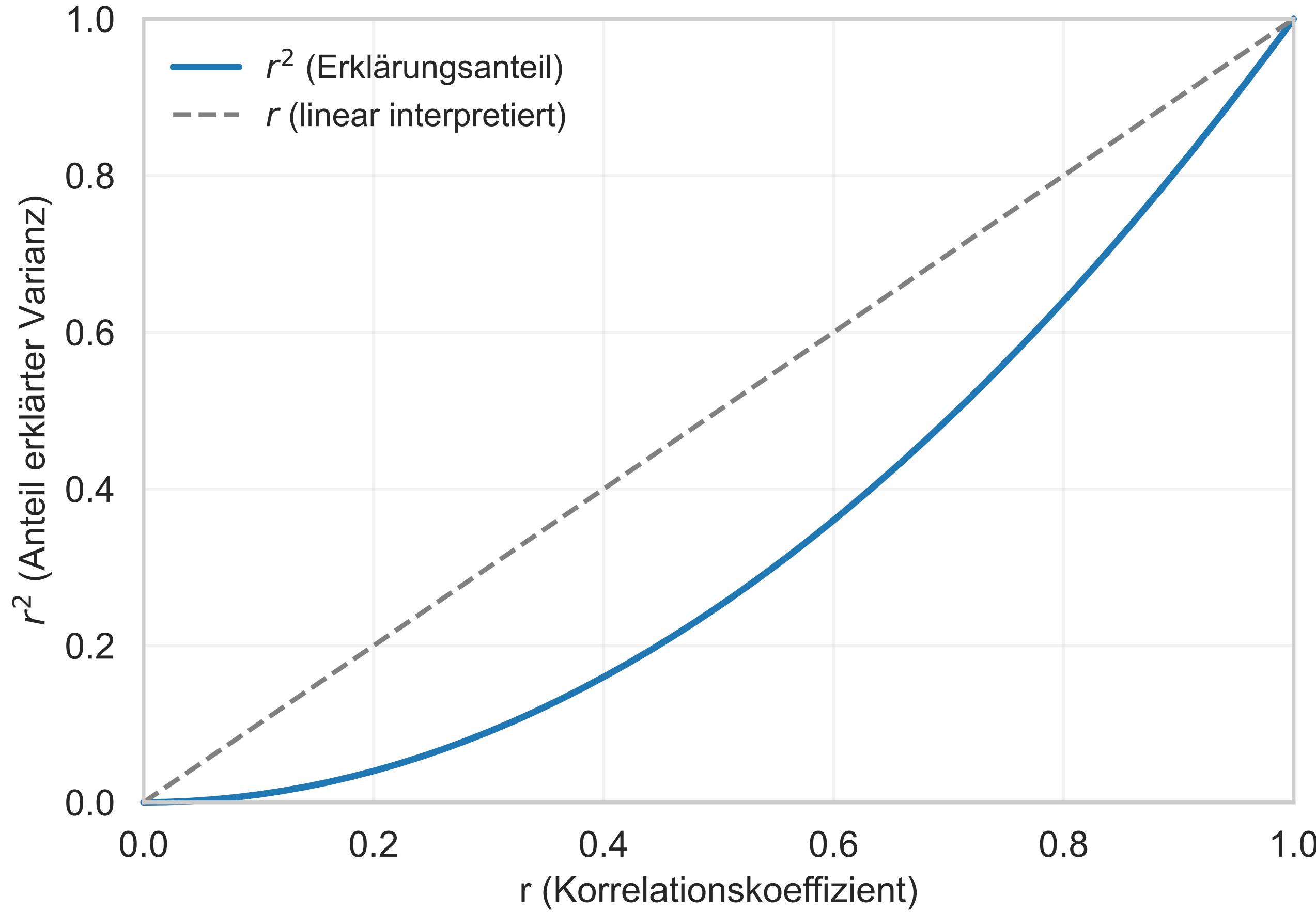


Table 1: Interpretation des Korrelationskoeffizienten r

r-Wert	Richtung	Stärke	Interpretation (sprachlich)
–1.0 bis –0.9	negativ	sehr stark	Fast perfekter gegenläufiger Zusammenhang
–0.9 bis –0.7	negativ	stark	Klarer Trend: je mehr X , desto weniger Y
–0.7 bis –0.4	negativ	mittel	Merklicher, aber nicht dominanter Zusammenhang
–0.4 bis –0.2	negativ	schwach	Leichter gegenläufiger Trend, viel Rauschen
–0.2 bis 0.2	–	keiner	Praktisch kein linearer Zusammenhang
0.2 bis 0.4	positiv	schwach	Leichter gemeinsamer Trend
0.4 bis 0.7	positiv	mittel	Stabiler, aber nicht perfekter Zusammenhang
0.7 bis 0.9	positiv	stark	Klarer Trend: X und Y steigen gemeinsam
0.9 bis 1.0	positiv	sehr stark	Fast perfekter Gleichlauf

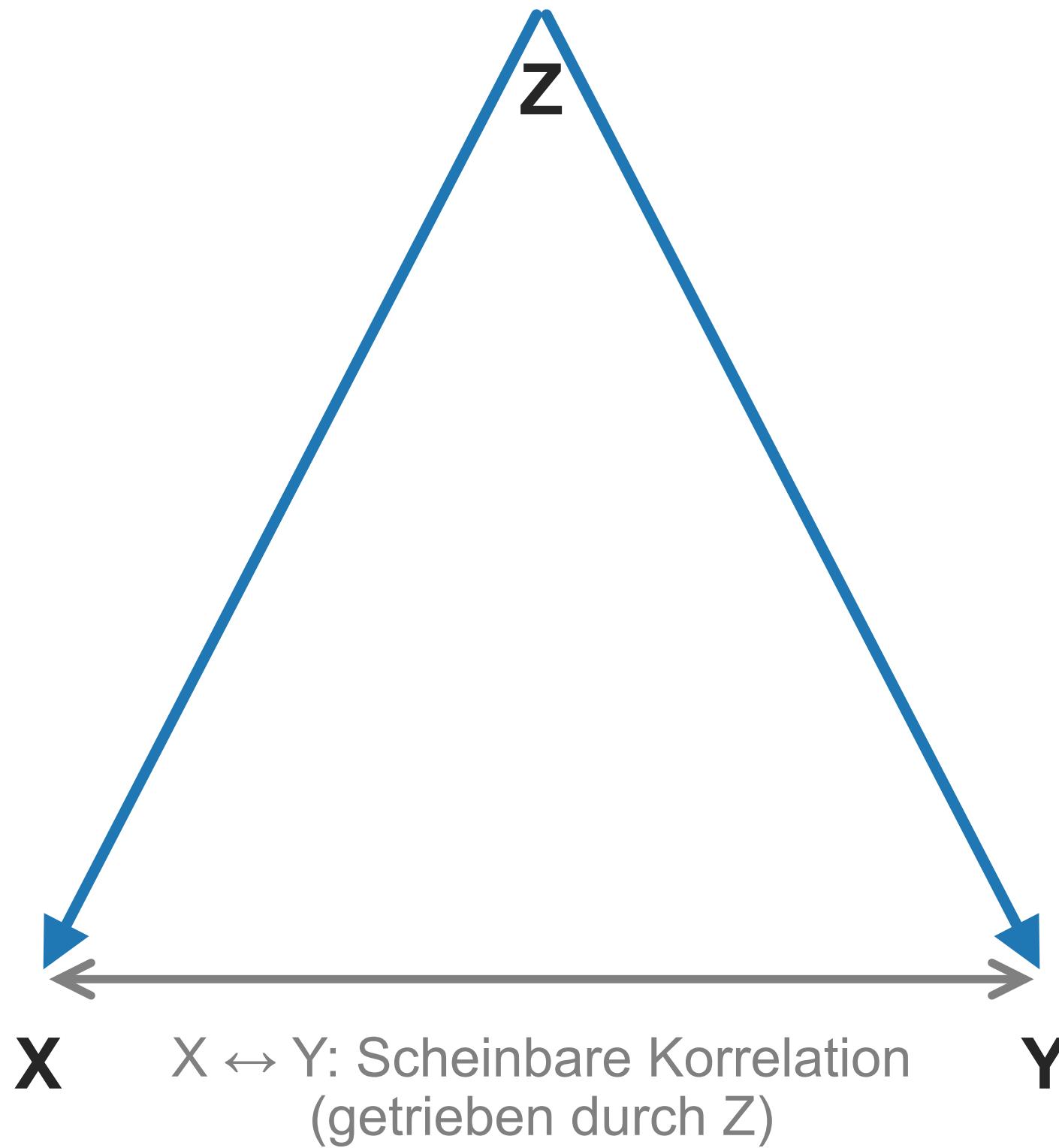
Korrelation vs. Kausalität

HAUPTREGEL: Korrelation \neq Beweis für Ursache (Kausalität).

- Der Übeltäter: **Drittvariablen** (Konfundierung) und reiner **Zufall**.
- Beispiel-Klassiker: Eisverkauf \leftrightarrow Ertrinken. (Der gemeinsame Treiber ist die Temperatur Z).
- Kausale Schlüsse benötigen **experimentelle Kontrolle**: Korrelation reicht dafür nicht.

Mini-Check: Wann kannst du aus Korrelation auf Kausalität schliessen?

Z = Drittvariable (z. B. Temperatur)



Erste mathematische Intuition: Abweichung

Die Idee: Wenn X und Y gemeinsam über/unter ihrem Mittelwert (\bar{x} , \bar{y}) liegen \Rightarrow positives Produkt.

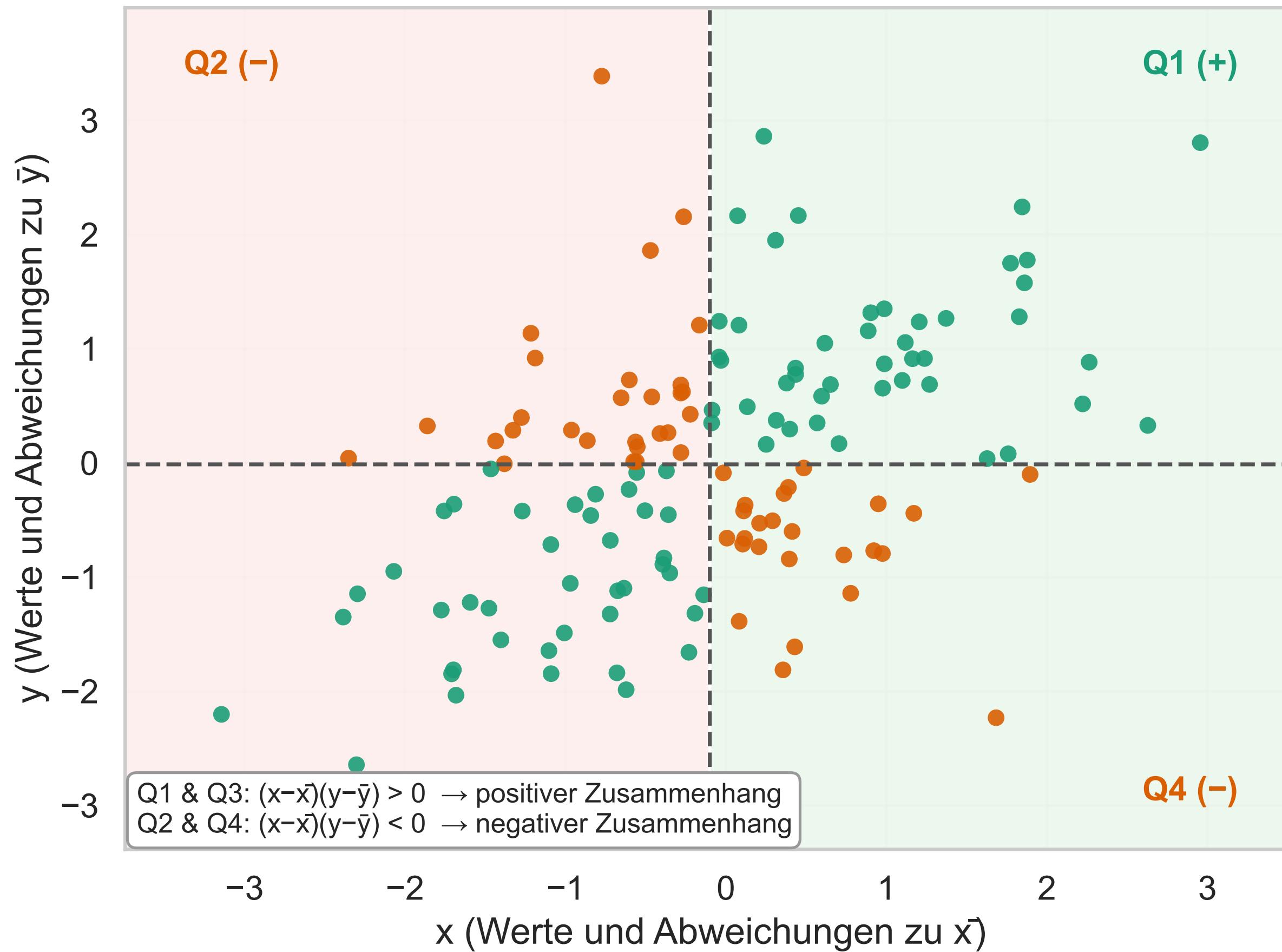
Das Produkt der Abweichungen misst dieses gemeinsame Verhalten.

- Quadrant I & III (Positives Produkt) \Rightarrow Positiver Zusammenhang.
- Quadrant II & IV (Negatives Produkt) \Rightarrow Negativer Zusammenhang.

Das ist die Grundlage der Kovarianz.

Mini-Check: Was passiert, wenn eine Variable konstant ist?

Produkt der Abweichungen: Quadranten und Vorzeichen



Beispiel mit Zahlen

Datensatz: $X = [1, 2, 3, 4]$.

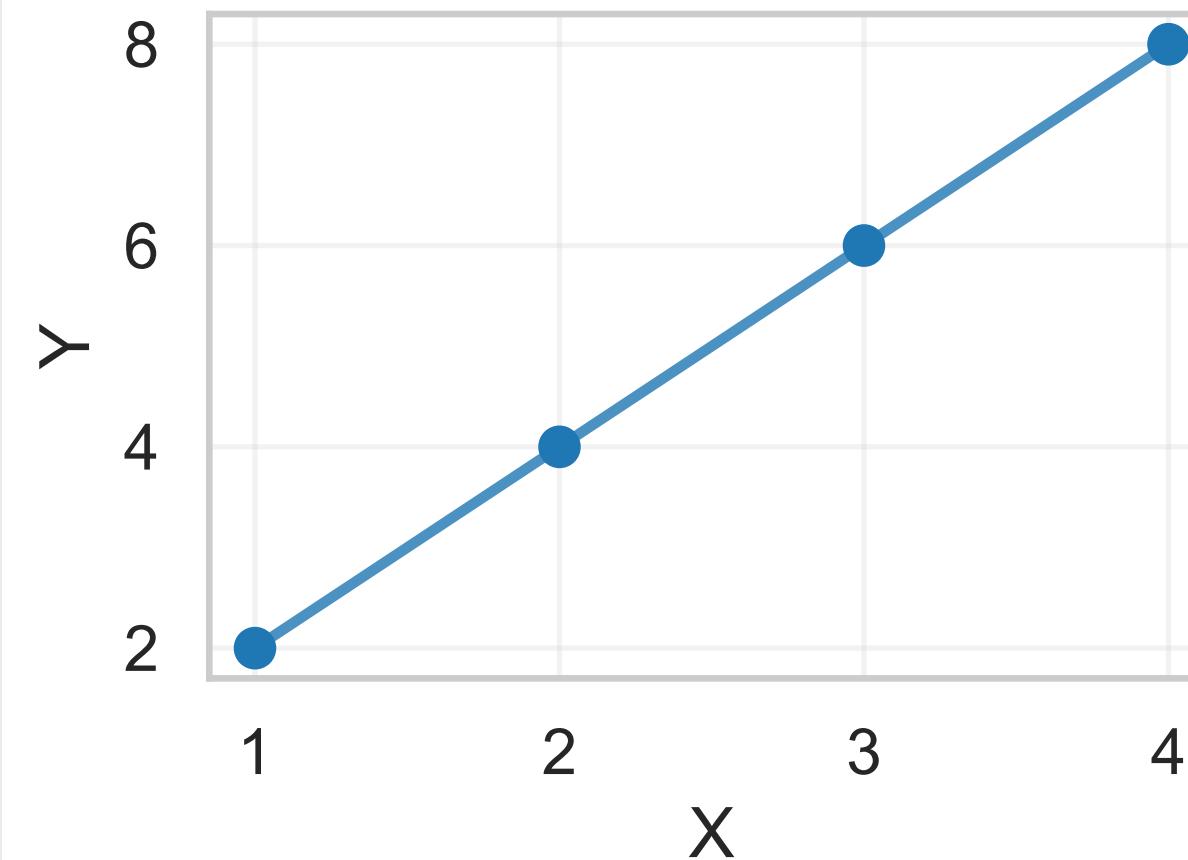
- Fall 1: $Y = [2, 4, 6, 8] \Rightarrow r = +1$ (Perfekt positiv).
- Fall 2: $Y = [8, 6, 4, 2] \Rightarrow r = -1$ (Perfekt negativ).
- Fall 3: $Y = [4, 3, 6, 5] \Rightarrow r \approx 0$ (Kein linearer Zusammenhang).

Korrelation entsteht aus **koordinierter Abweichung** vom Mittelwert.

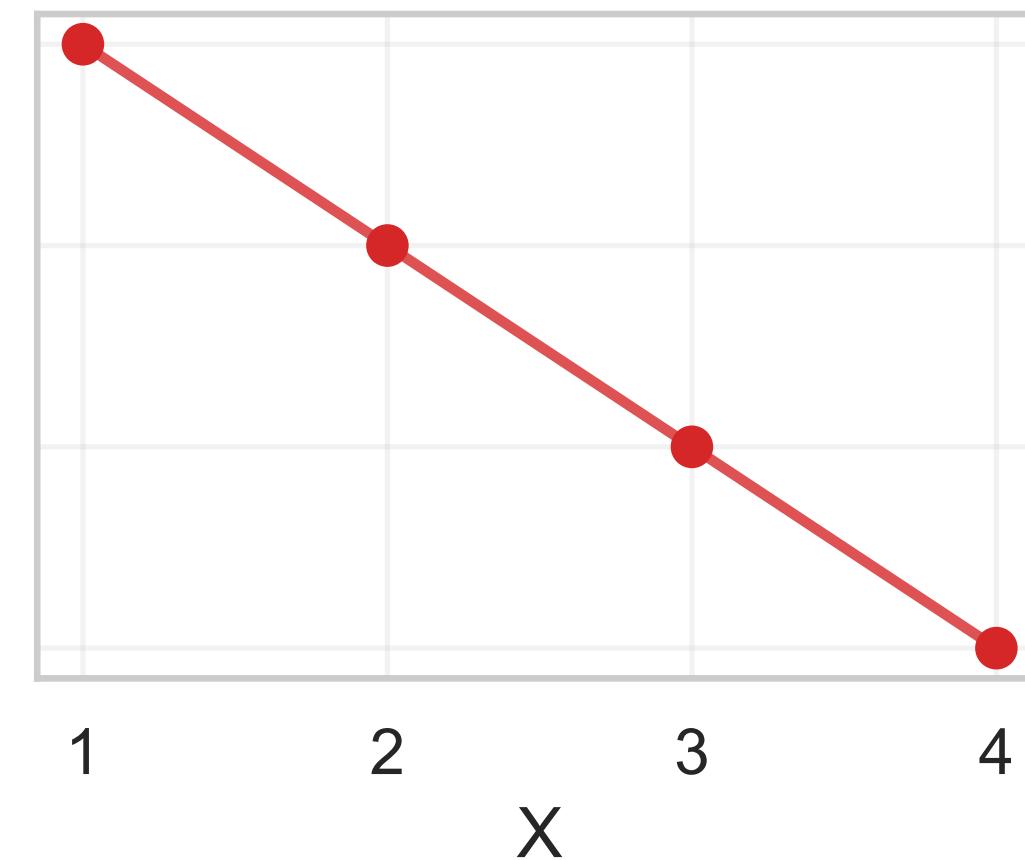
Mini-Check: Welche Beziehung zeigt $r = -1$?

Beispiel mit Zahlen – Korrelation als koordinierte Abweichung

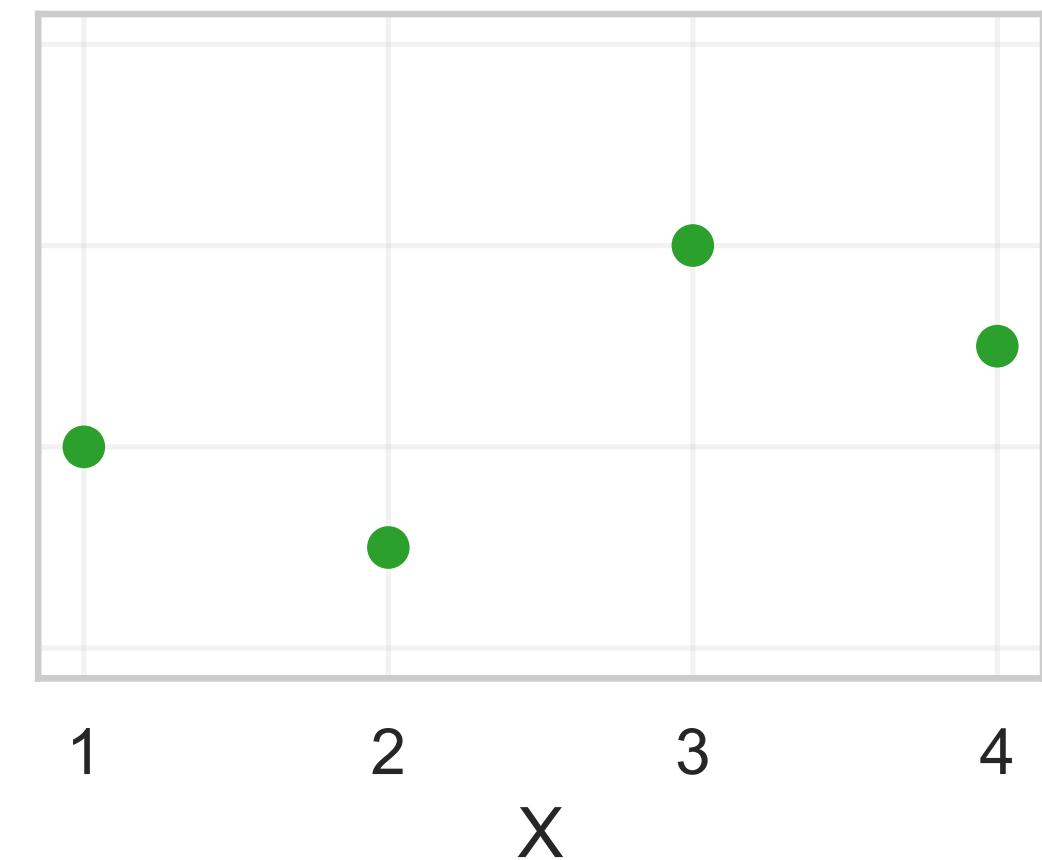
Fall 1: $r = +1.00$



Fall 2: $r = -1.00$



Fall 3: $r \approx 0.60$



Interaktive Mini-Frage: Visuelle Schätzung

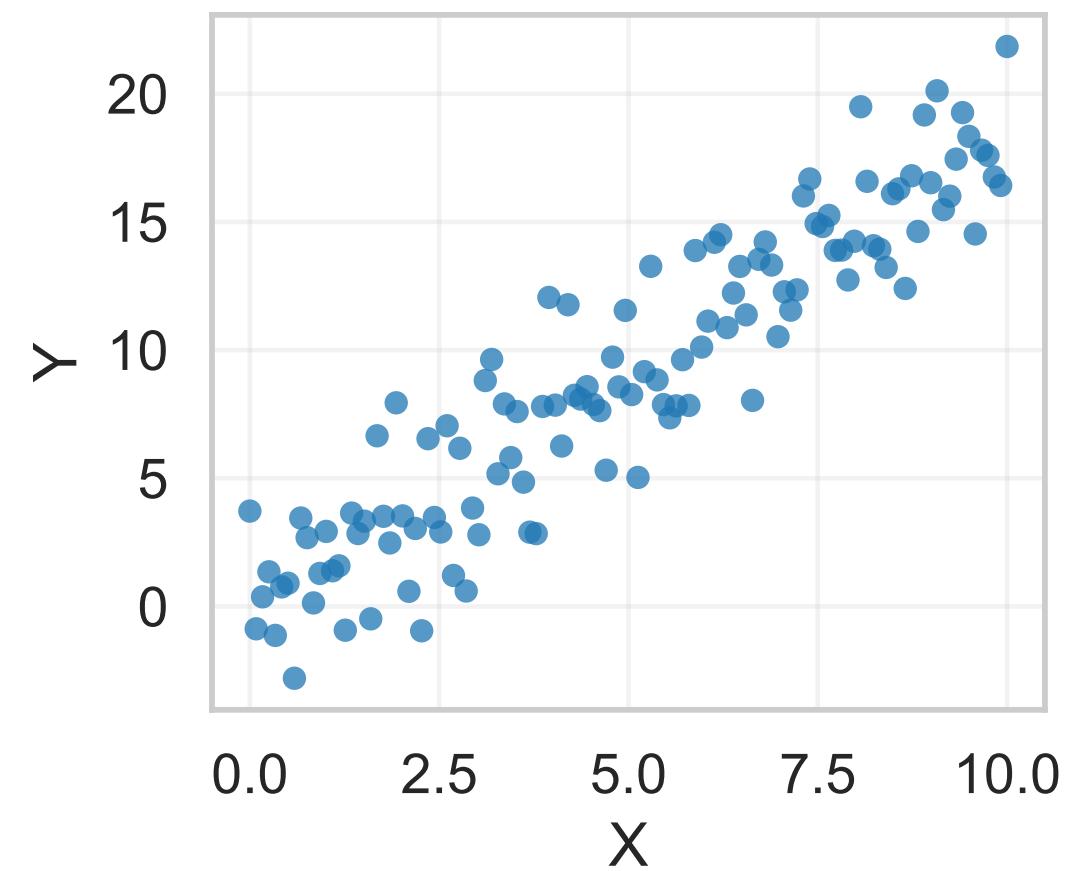
Stell dir 3 anonyme Scatterplots (A , B , C) vor.

- Frage an dich (1 min):
 1. Welcher Plot hat die **stärkste** Korrelation (Betrag)?
 2. Welcher hat die **negativste** Korrelation?
- Ziel: Visuelle Einschätzung üben: aber **sei kritisch**.

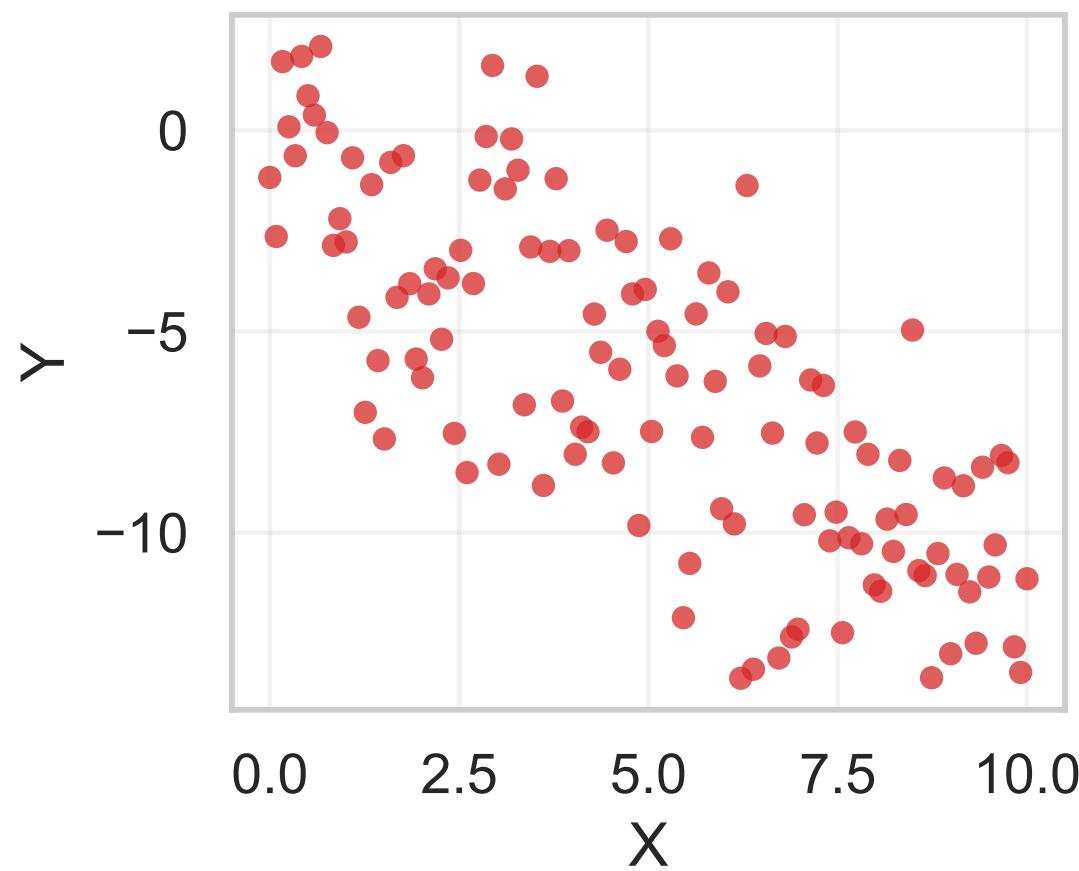
Mini-Check: Einschätzen

Interaktive Mini-Frage: Visuelle Schätzung (A/B/C)

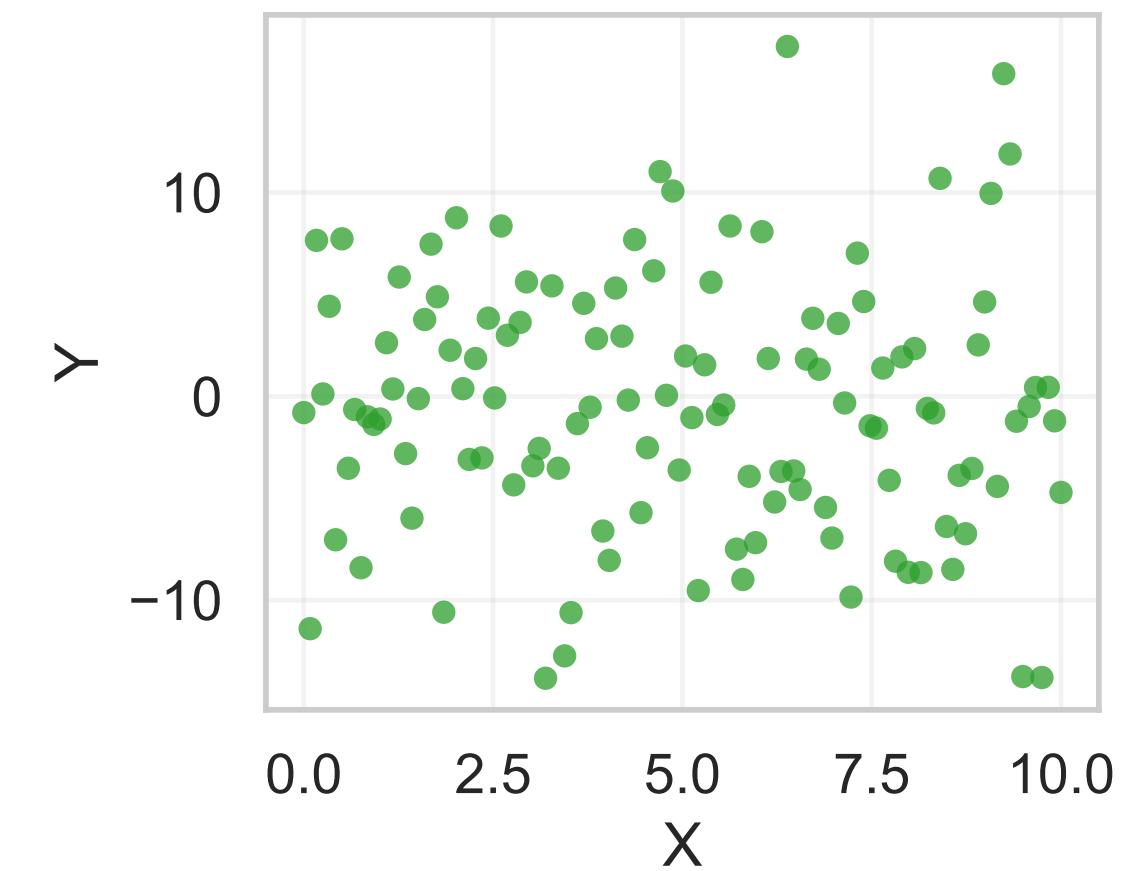
Plot A



Plot B



Plot C



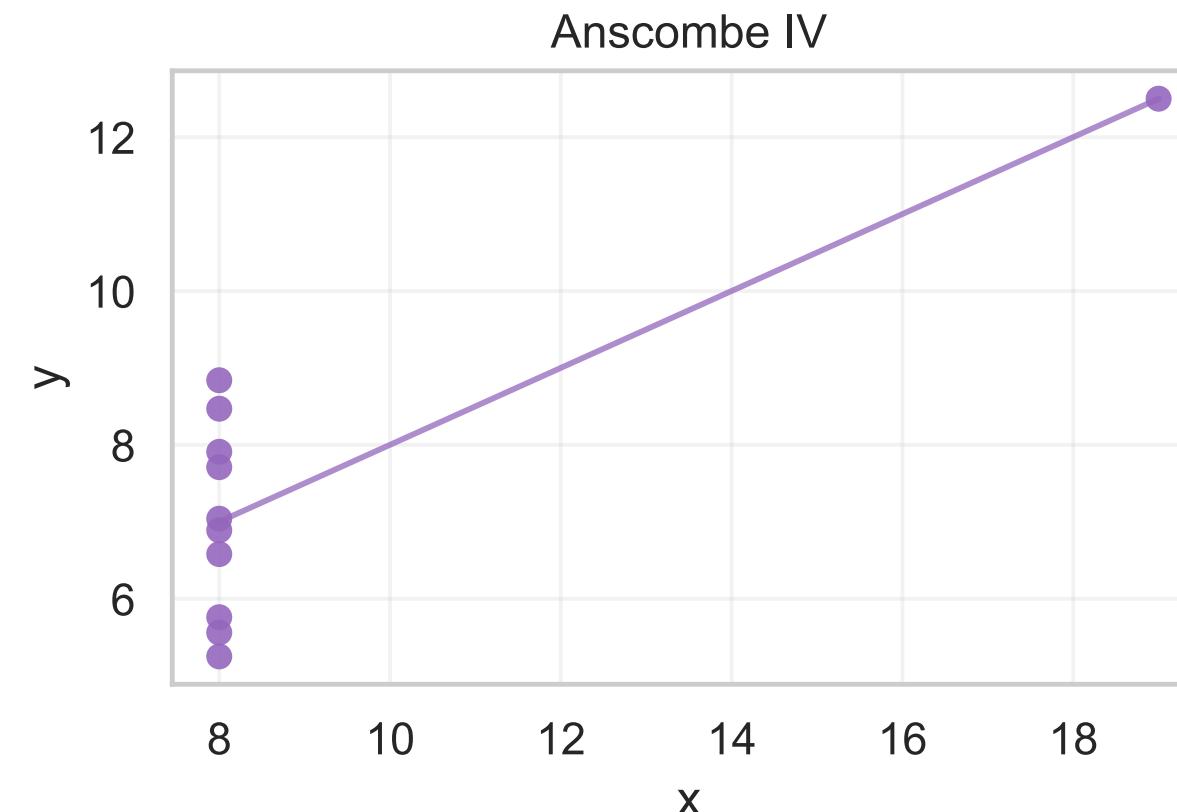
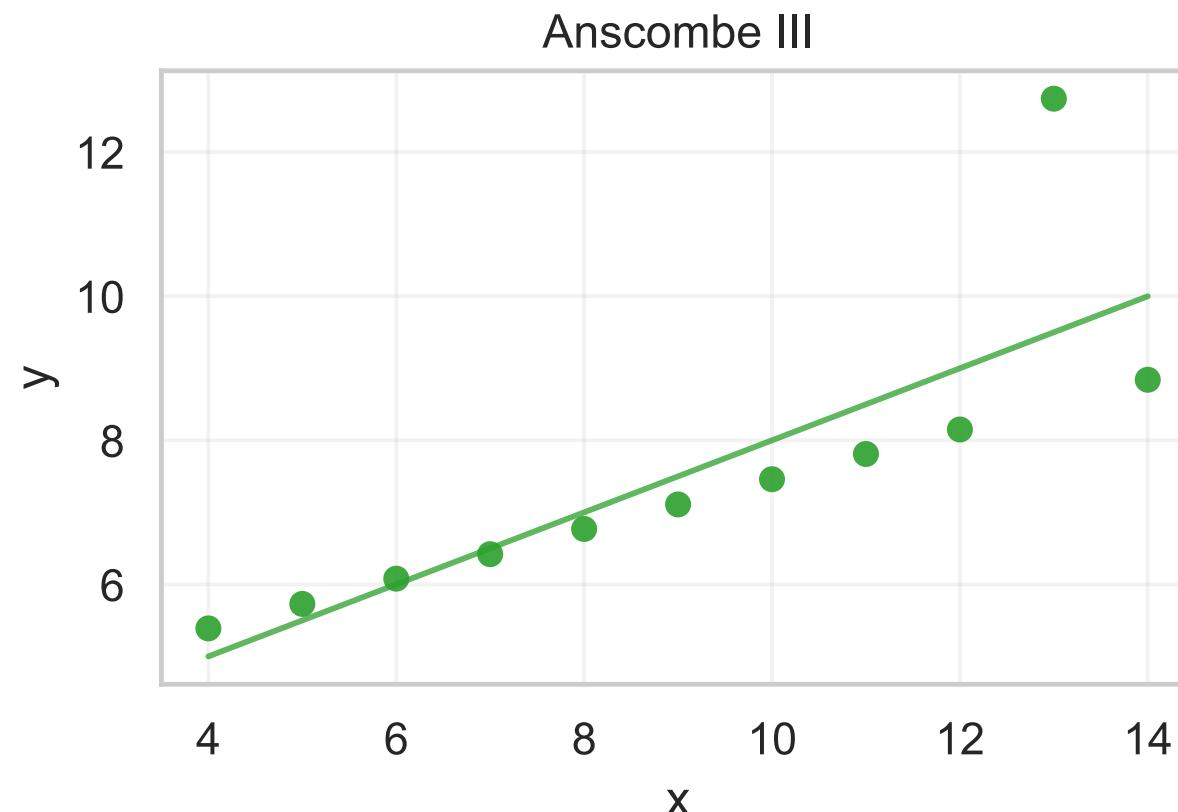
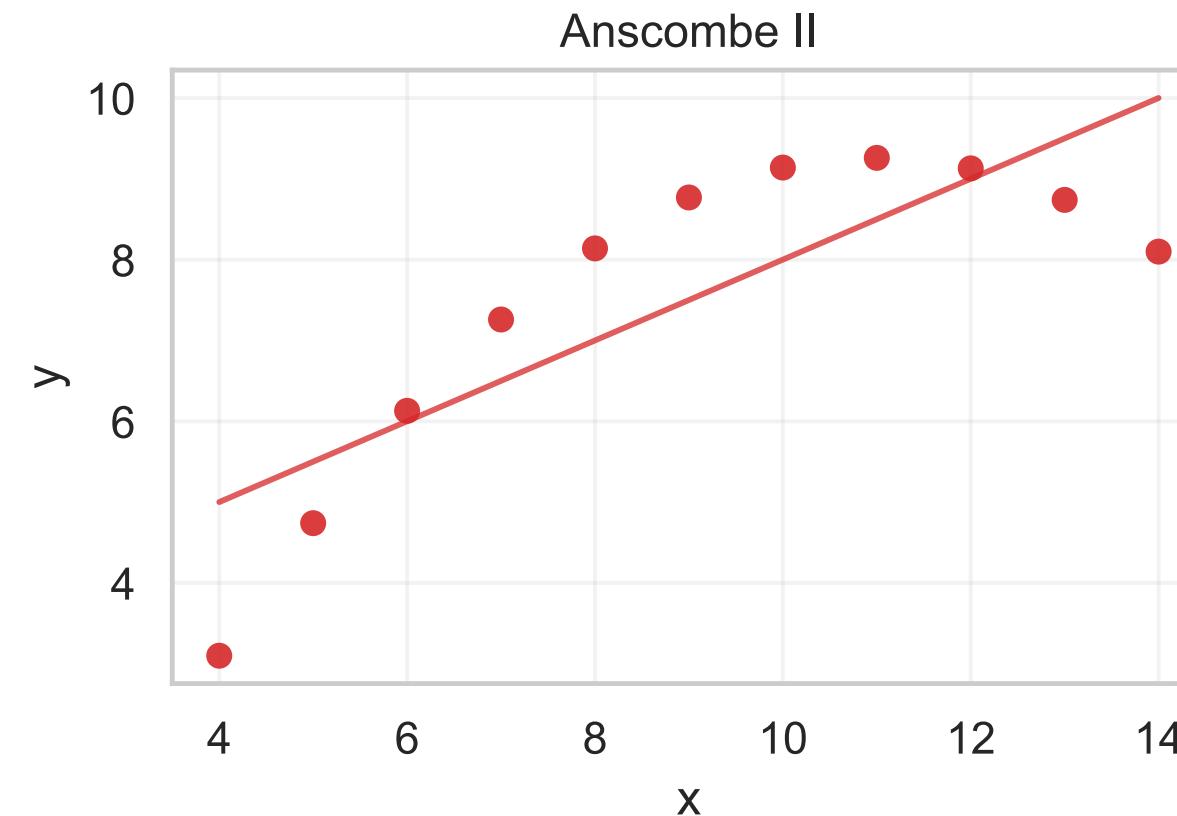
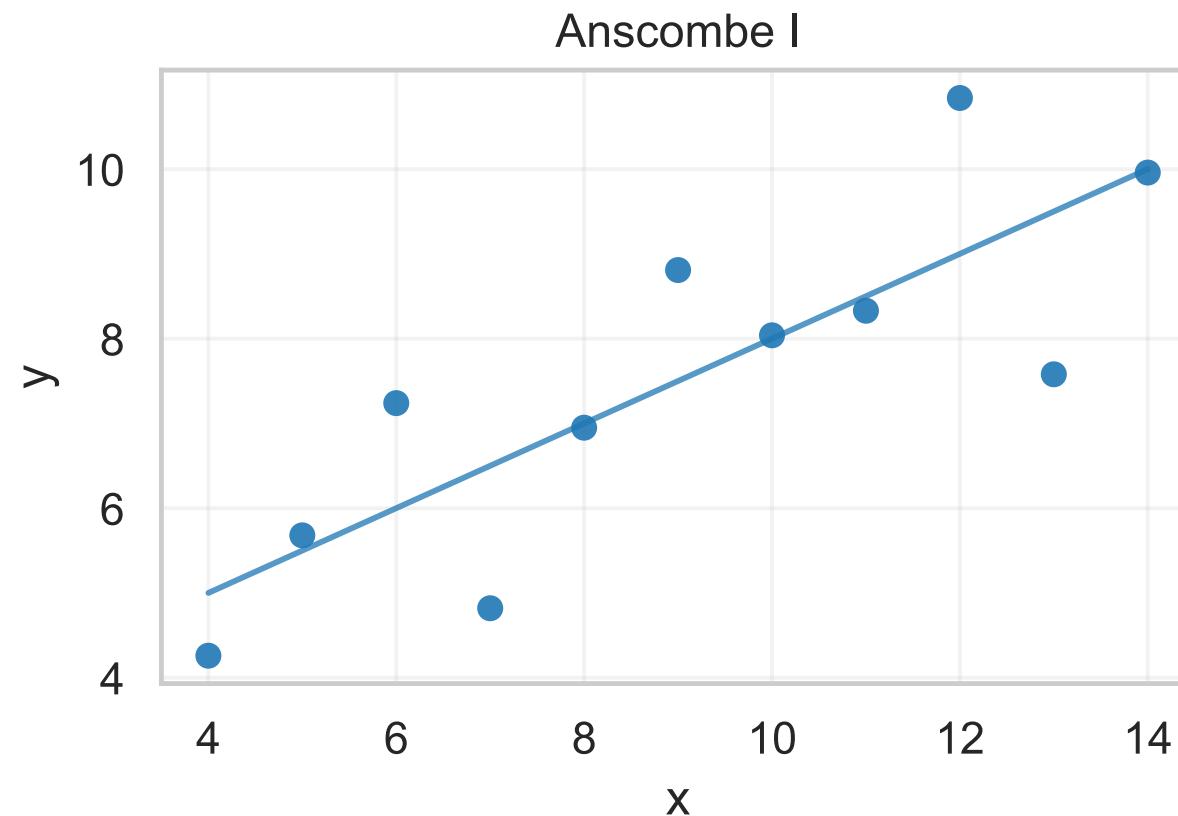
Anscombe's Quartett

Zeige Anscombe's Quartett (4 Datensätze mit identischen deskriptiven Statistiken).

- Alle 4 Datensätze haben:
 - Gleicher Korrelationskoeffizient $r = 0.816$.
 - Gleicher Mittelwert von X und $Y \Rightarrow \bar{x} = 9 \quad \bar{y} = 7.5$
 - **Fazit:** Immer plotten – verlass dich **nie** nur auf r .

Mini-Check: Warum ist r gleich, obwohl die Form der Punkte verschieden ist?

Anscombe's Quartett – gleiche Statistik, unterschiedliche Struktur



Key Takeaways: Korrelation, Kausalität & Form

- **Zusammenhangs-Check:** Korrelation liefert die erste, quantitative Antwort auf die Frage: «Ändert sich Y mit X ?»
 - **Richtung & Stärke:** Der Koeffizient r liegt zwischen $[-1, 1]$. Vorzeichen zeigt die Richtung, der Betrag die Stärke.
 - **Form matters:** Unterscheide zwischen **linearen** ($y \approx a + bx$) und **monotonen** Zusammenhängen.
 - **Goldene Regel: Korrelation \neq Kausalität.** Scheinkorrelationen entstehen oft durch Drittvariablen.
 - **Visualisierungspflicht:** Nur der Scatterplot zeigt die tatsächliche Struktur (Anscombe's Quartett!).
- Always plot your data!**
- Francis John Anscombe 1973

Kovarianz & Pearson-Korrelation

Von der Idee zur Formel

Kovarianz misst, wie stark X und Y gemeinsam von ihren Mittelwerten abweichen.

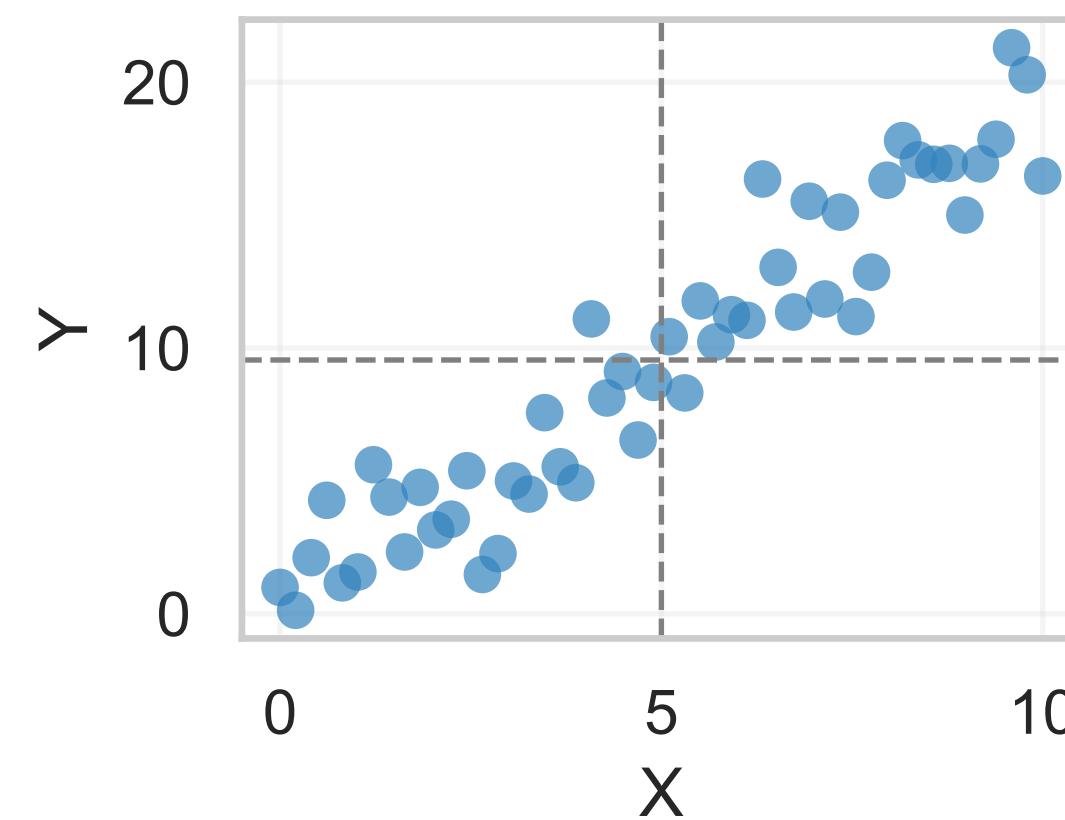
$$Cov(X, Y) = \frac{1}{n - 1} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- Positiv, wenn beide gemeinsam über/unter Mittelwert liegen.
- Negativ, wenn gegensätzlich.
- **Wichtig:** Keine Normierung \Rightarrow abhängig von Einheiten.

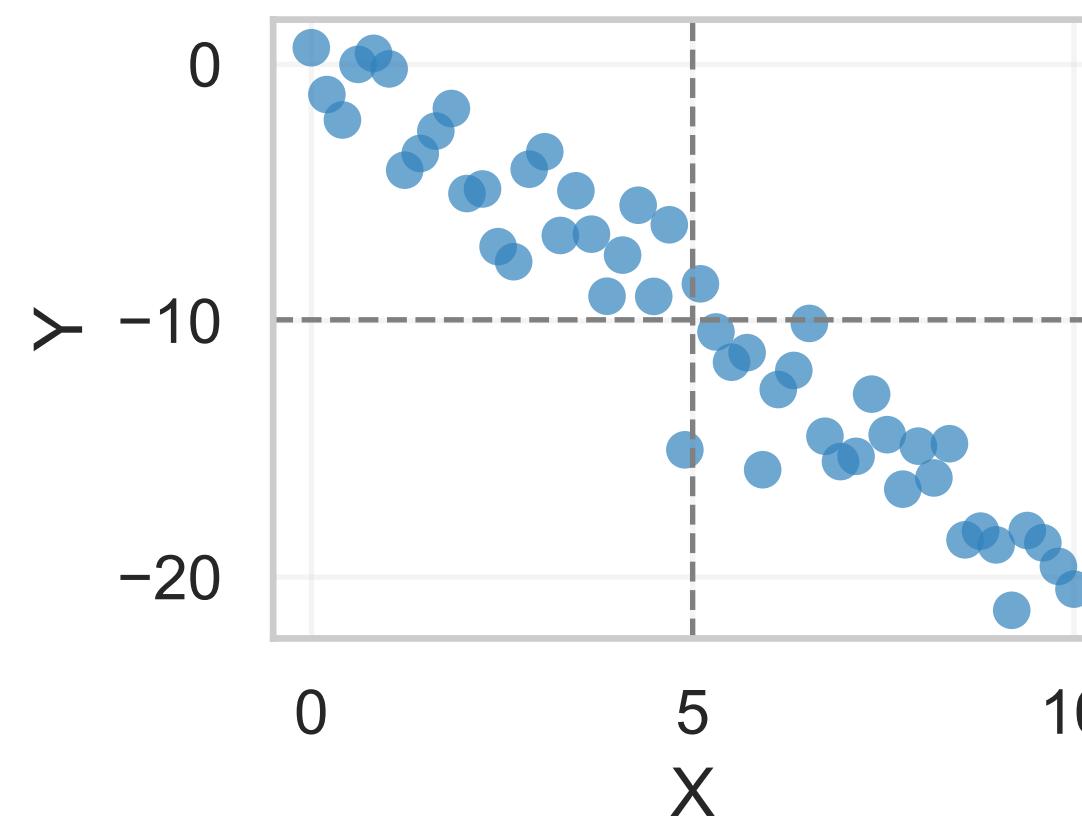
Mini-Check: Was passiert mit $Cov(X, Y)$, wenn Y konstant ist?

Kovarianz – gemeinsame Abweichung vom Mittelwert

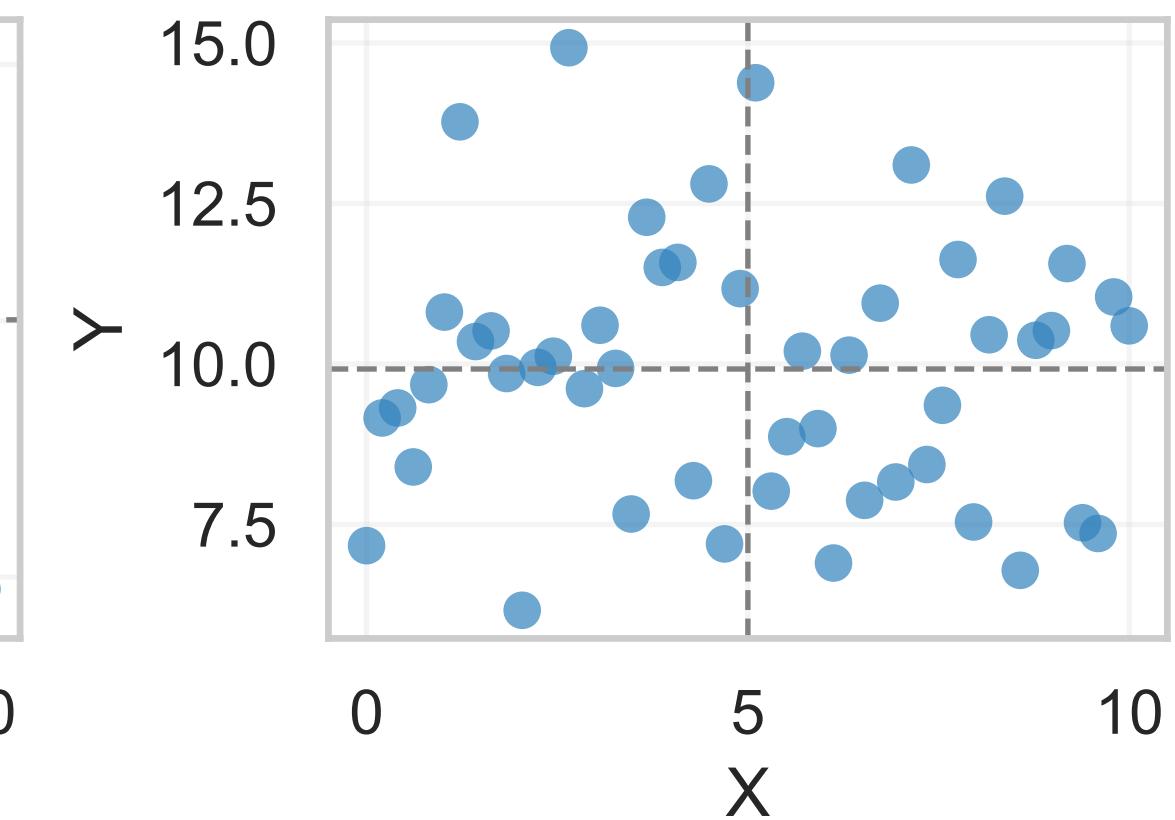
Positive Kovarianz



Negative Kovarianz



Keine Kovarianz



Numerisches Beispiel: Kovarianz

Zahlenbeispiel macht gemeinsame Streuung sichtbar.

- $X = [1, 2, 3, 4]$, $Y = [2, 4, 6, 8]$.
- Mittelwerte: $\bar{x} = 2.5$, $\bar{y} = 5$.
- Produkte der Abweichungen: $(+ \times + = +)$, $Cov > 0$.
- Ergebnis: $Cov \approx 3.33$.

Mini-Check: Was wäre Cov , wenn $Y = [8, 6, 4, 2]$?

Table 1: Numerisches Beispiel: Kovarianz

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	Produkt
1	1	2	-1.5	-3	+4.5
2	2	4	-0.5	-1	+0.5
3	3	6	+0.5	+1	+0.5
4	4	8	+1.5	+3	+4.5
Summe					10.0

$$\text{Cov}(X, Y) = \frac{10}{3} \approx 3.33$$

Problem: Skalenabhängigkeit

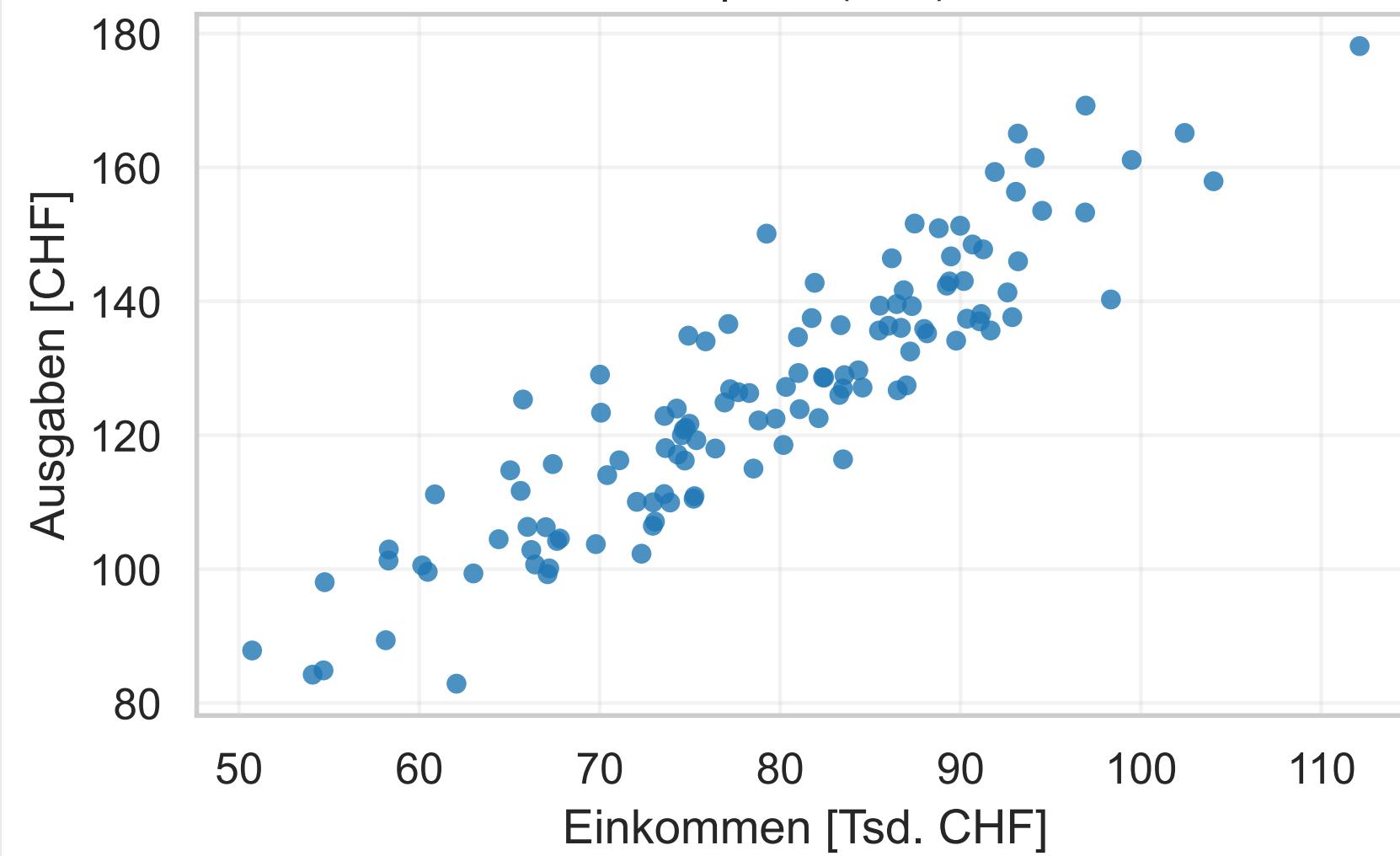
Kovarianz hängt von Mess-Einheiten ab. Das ist ihr Todesurteil für Vergleiche.

- Beispiel: $Cov(\text{Grösse}, \text{Gewicht}) = +200 \Rightarrow$ nicht vergleichbar mit $Cov(\text{Einkommen}, \text{Zufriedenheit})$.
- Die Lösung: Standardisierung auf Einheitsstreuung \Rightarrow Korrelation (r).
- Motivation für Pearson r .

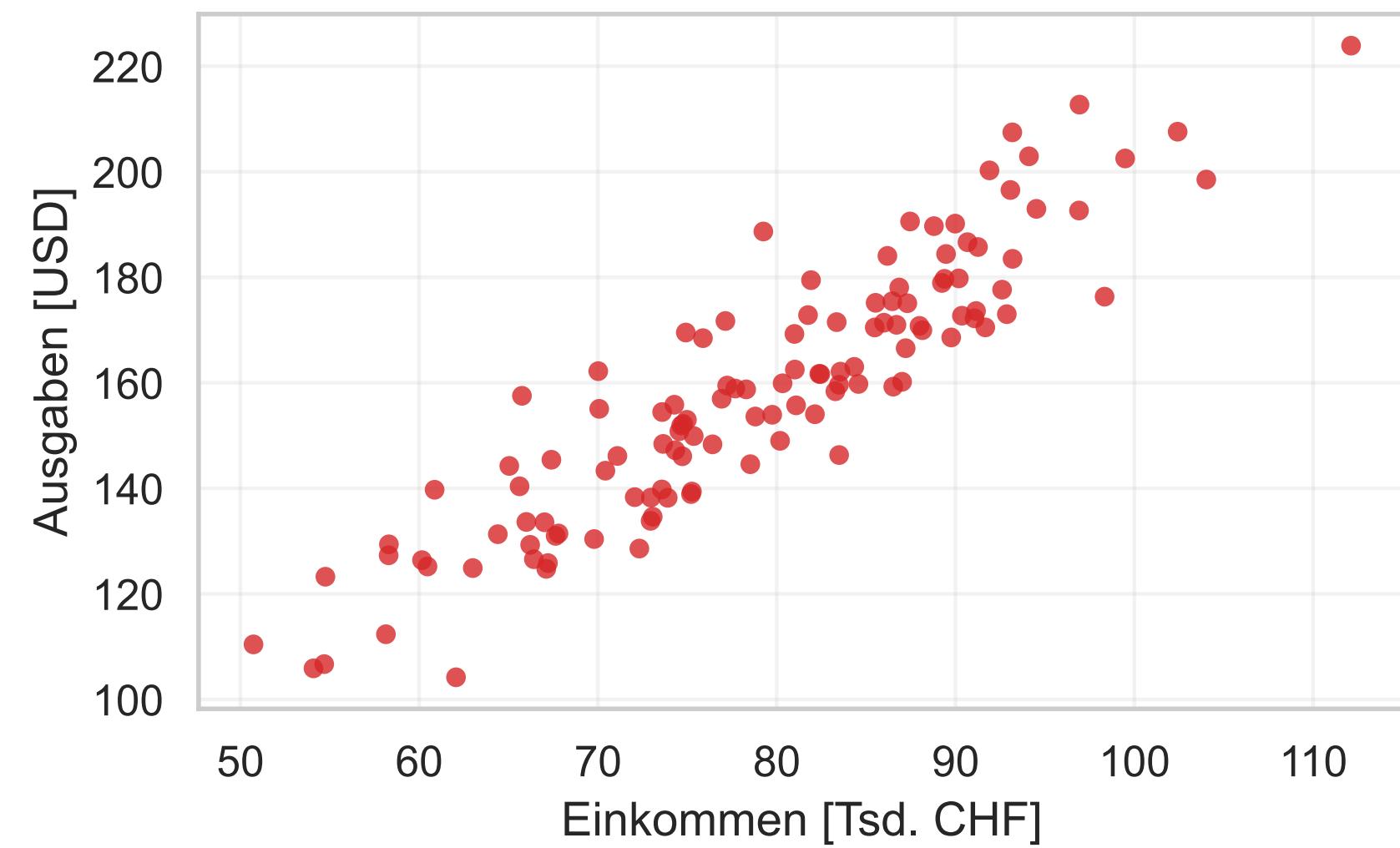
Mini-Check: Wie ändert sich $Cov(X, Y)$, wenn Y in \$ statt CHF gemessen wird?

Skalenabhängigkeit der Kovarianz: CHF → USD

CHF-Skala | $\text{Cov}(X, Y) = 209.70$



USD-Skala ($\times 1.2569$) | $\text{Cov}(X, Y) = 263.57$
 $\approx 1.2569 \times 209.70$



Definition: Pearson-Korrelation

Pearson r ist die standardisierte Kovarianz.

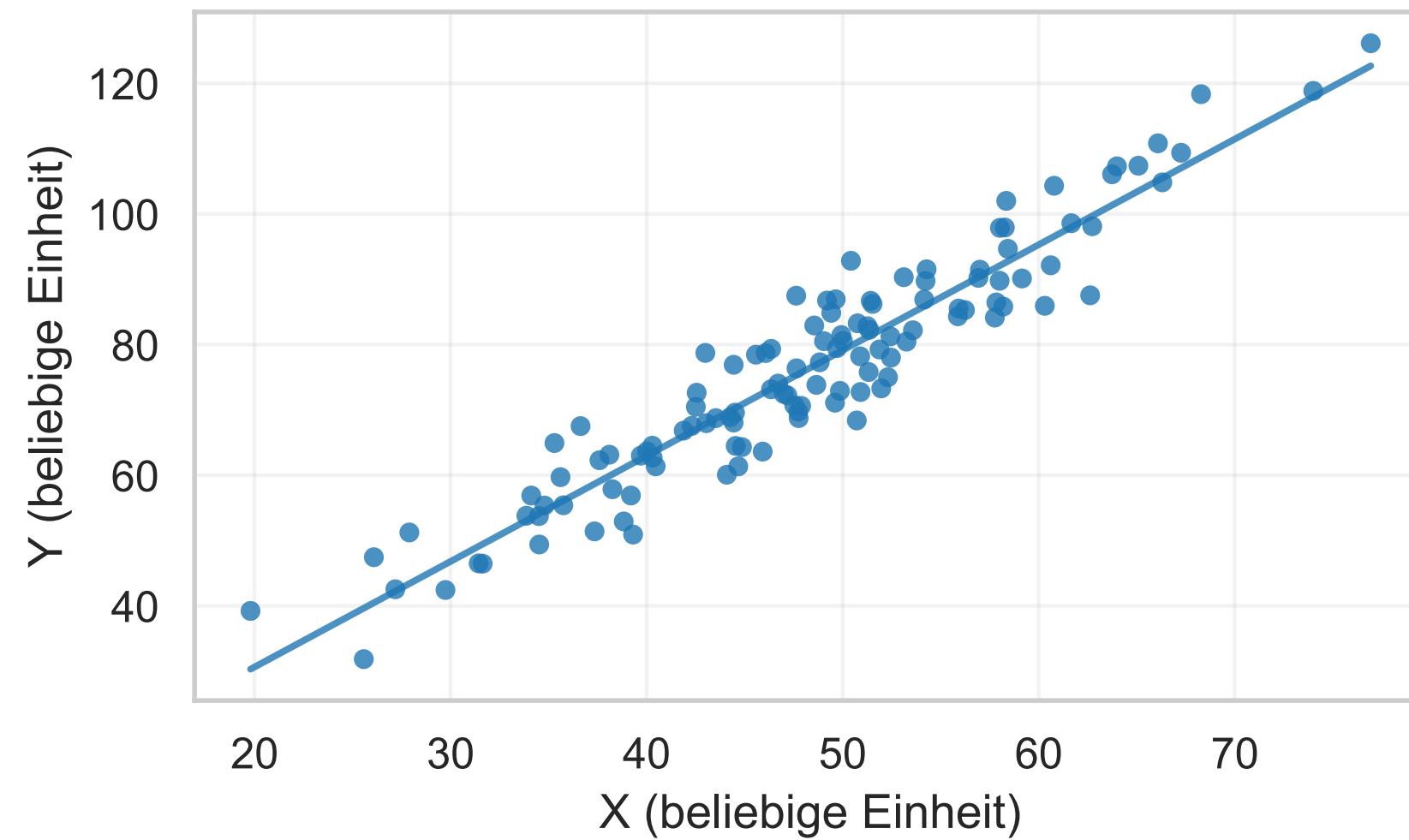
$$r_{xy} = \frac{\text{Cov}(X, Y)}{s_X s_Y}$$

- $r \in [-1, 1]$.
- Misst **linearen** Zusammenhang, symmetrisch ($r_{xy} = r_{yx}$).
- Unempfindlich gegen Skalierung ($X \times k$).

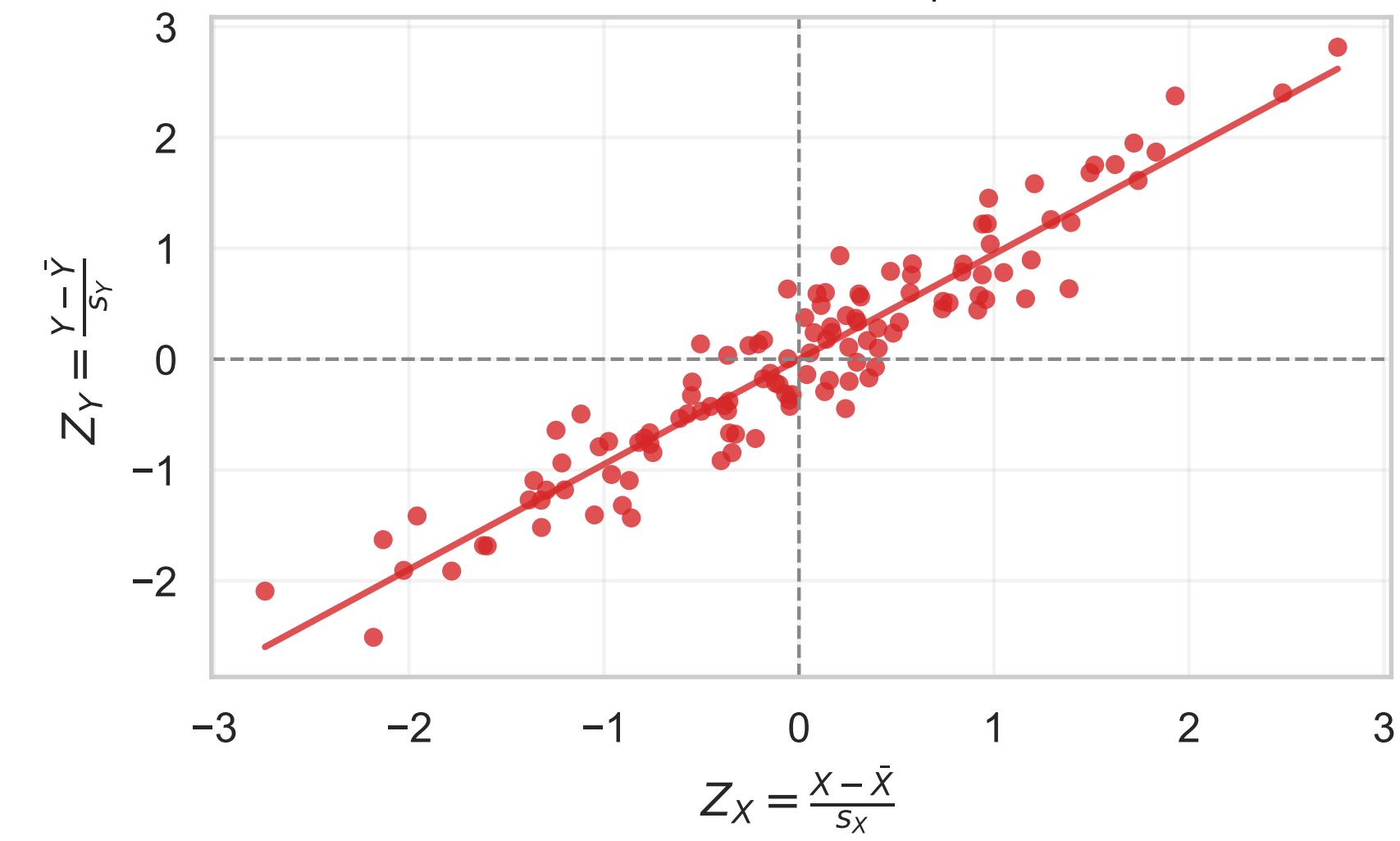
Mini-Check: Welche Einheit hat r ?

Pearson r ist die standardisierte Kovarianz: $r_{xy} = \frac{\text{Cov}(X, Y)}{s_X s_Y}$

Rohdaten (X, Y) | $r = 0.95$



Standardisiert (Z_X, Z_Y) | $r = 0.95$



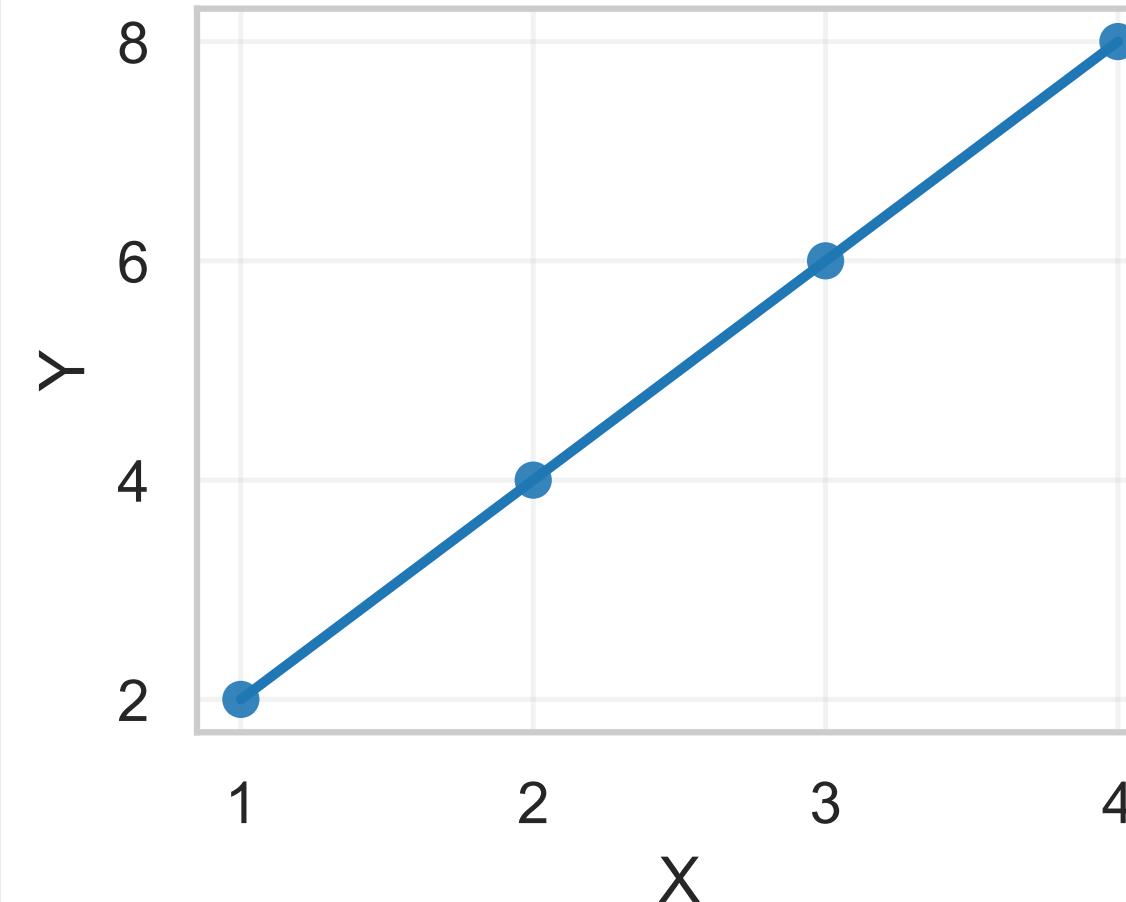
Beispiel: Berechnung von r

- $X = [1, 2, 3, 4], Y = [2, 4, 6, 8] \Rightarrow r = 1.0.$
- $X = [1, 2, 3, 4], Y = [8, 6, 4, 2] \Rightarrow r = -1.0.$
- $X = [1, 2, 3, 4], Y = [3, 3, 5, 5] \Rightarrow r \approx 0.73.$

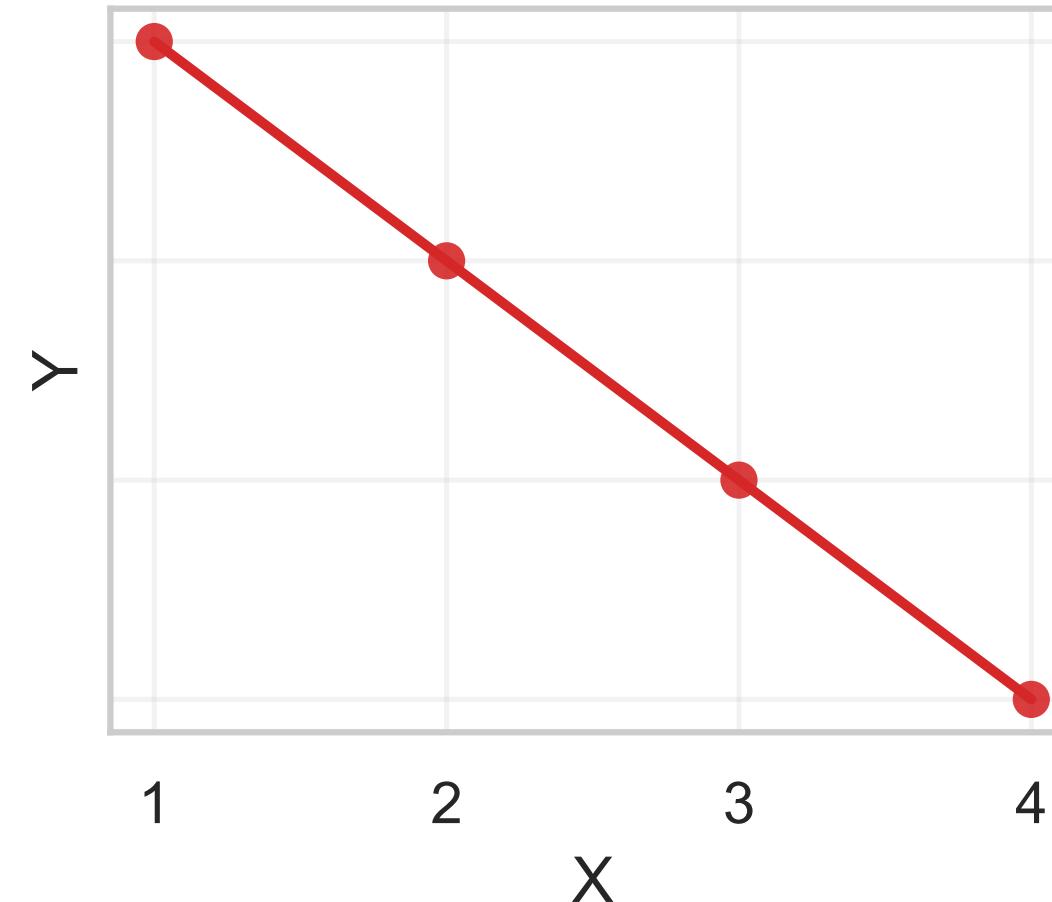
Mini-Check: Was bedeutet $r = 0.73$ im Kontext von X, Y ?

Beispiel: Berechnung von r (linearer Zusammenhang)

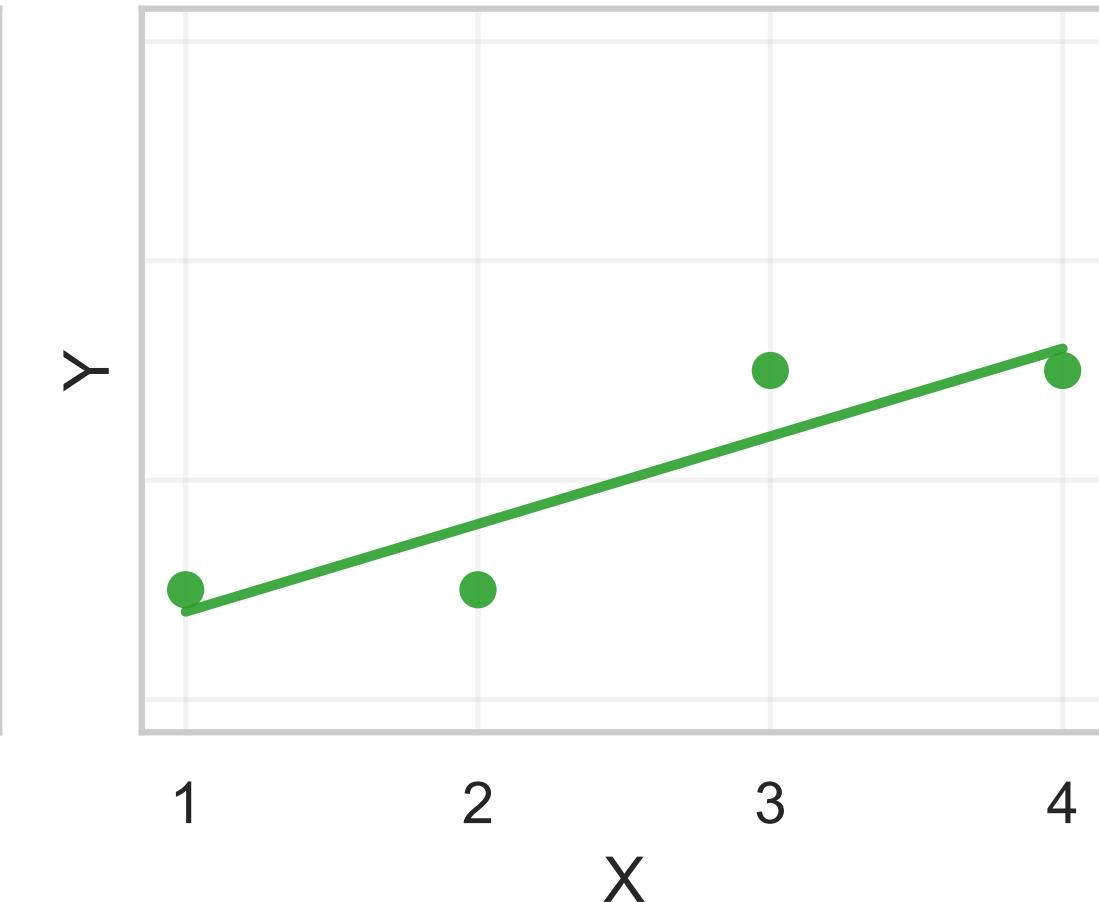
Fall 1: $r = 1.00$



Fall 2: $r = -1.00$



Fall 3: $r = 0.89$



Pearson in Python (1)

```
import numpy as np, scipy.stats as st
# Perfekt linearer Datensatz
x = np.arange(1,6)
y = np.array([2,4,6,8,10])
r, p = st.pearsonr(x,y)
# Ausgabe: (Korrelationskoeffizient, p-Wert)
print(f"r: {r:.2f}, p: {p:.2e}")
```

r: 1.00, p: 0.00e+00

Pearson in Python (2)

```
r: 1.00, p: 0.00e+00
```

- Ausgabe: $r \approx 1.00$, $p \approx 0.0$.
- Ein perfekter linearer **Zusammenhang**: $r \approx 1.00$.
- p-Wert = **probability value** = Wahrscheinlichkeitswert.
- $p < 0.05$ bedeutet: Die Wahrscheinlichkeit, einen so starken Zusammenhang zufällig zu beobachten, wenn es in Wahrheit keinen echten Zusammenhang gibt, ist sehr klein (unter 5 %)

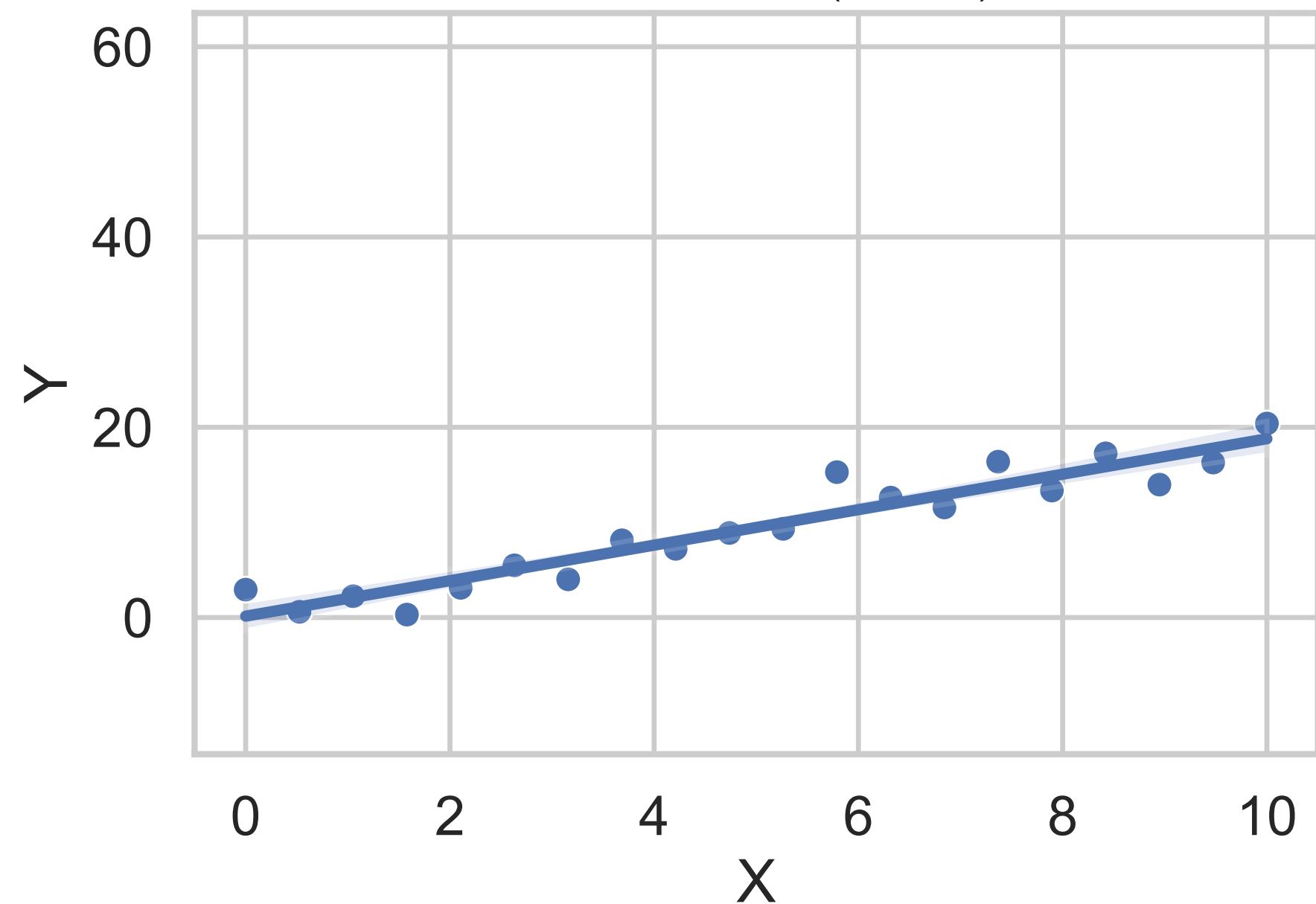
Einfluss von Aussreisern

Achtung, r ist sehr empfindlich. Ein einziger Aussreißer kann dein Ergebnis killen.

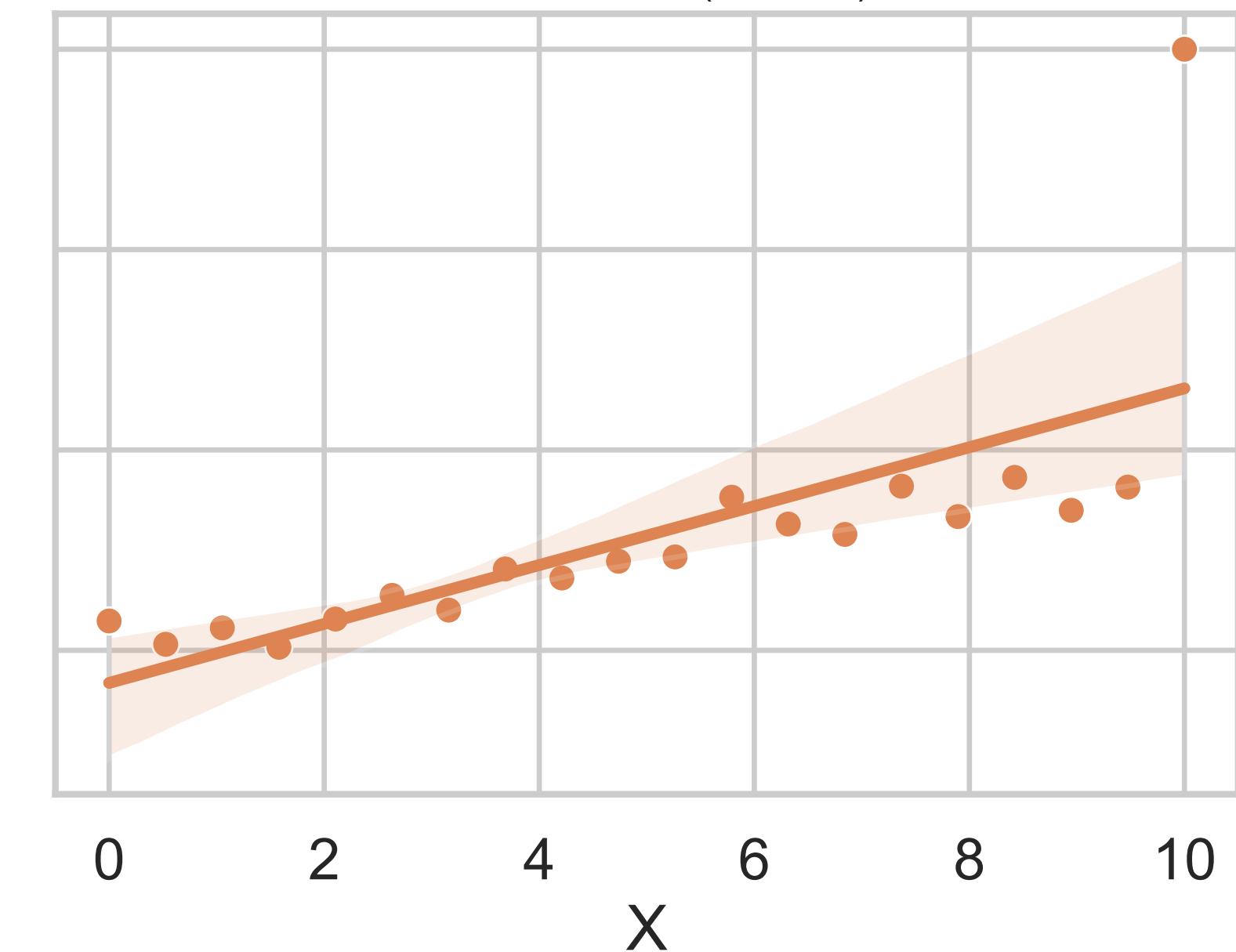
- Beispiel: r ohne Aussreißer = 0.9, mit Aussreißer = 0.2.
- Ursache: Quadratische Abweichungen verzerren Cov dramatisch.
- Die Alternative, wenn die Daten schräg sind: Rangbasierte Masse (Spearman).

Mini-Check: Wie würdest du Aussreißer prüfen, bevor du r rechnest?

Ohne Ausreißer ($r = 0.95$)



Mit Ausreißer ($r = 0.72$)



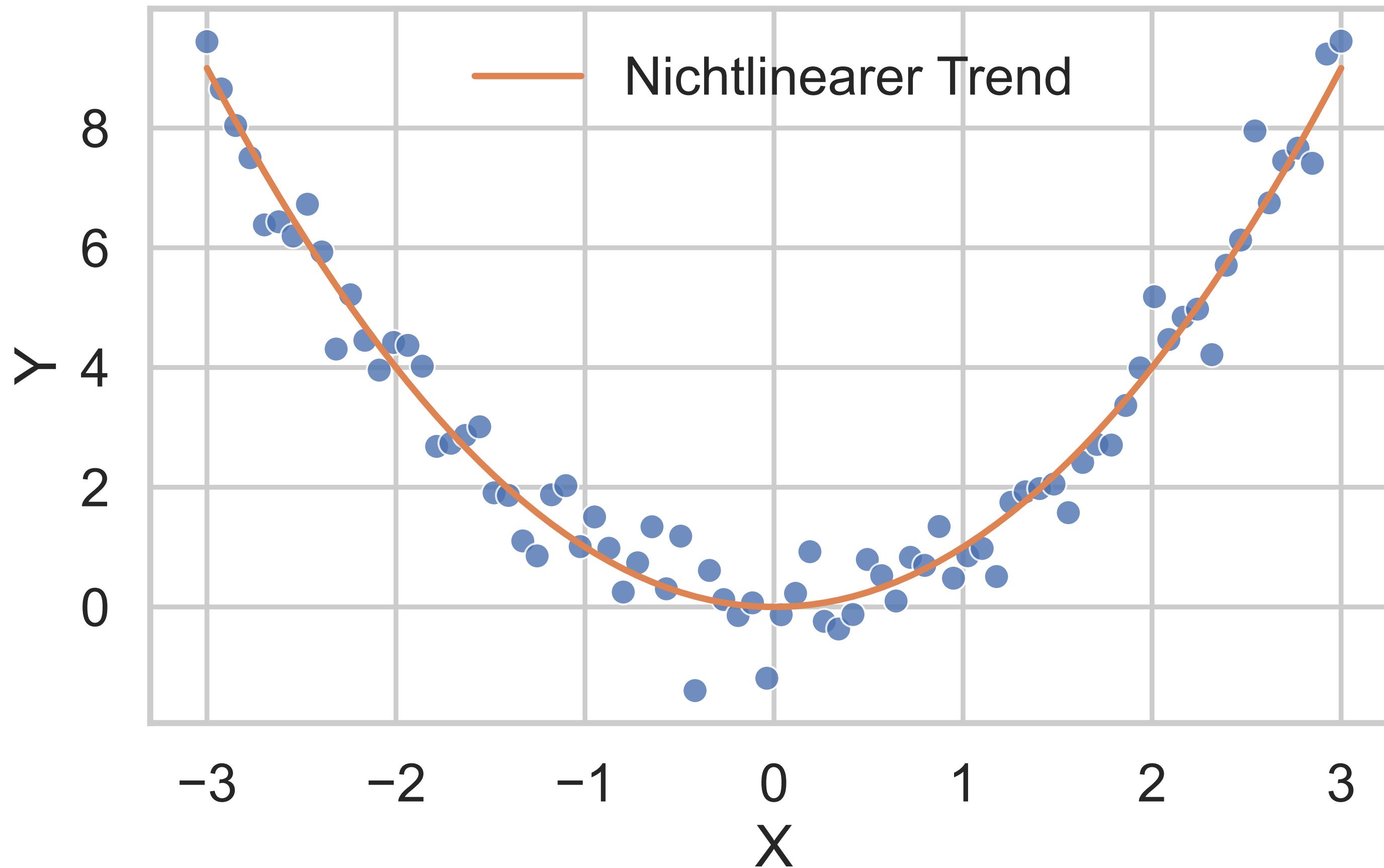
Voraussetzungen für Pearson

r ist ein gutes Werkzeug, wenn du die Spielregeln kennst.

- Variablen metrisch (Interval, Ratio). \Rightarrow nicht gut für Ordinal, Nominal
- **Lineare** Beziehung
- Keine starken Aussreisser.
- Annähernde Normalverteilung.

Mini-Check: Wie würdest du eine Nichtlinearität im Plot erkennen?

Nichtlinearer Zusammenhang ($r = 0.01$)



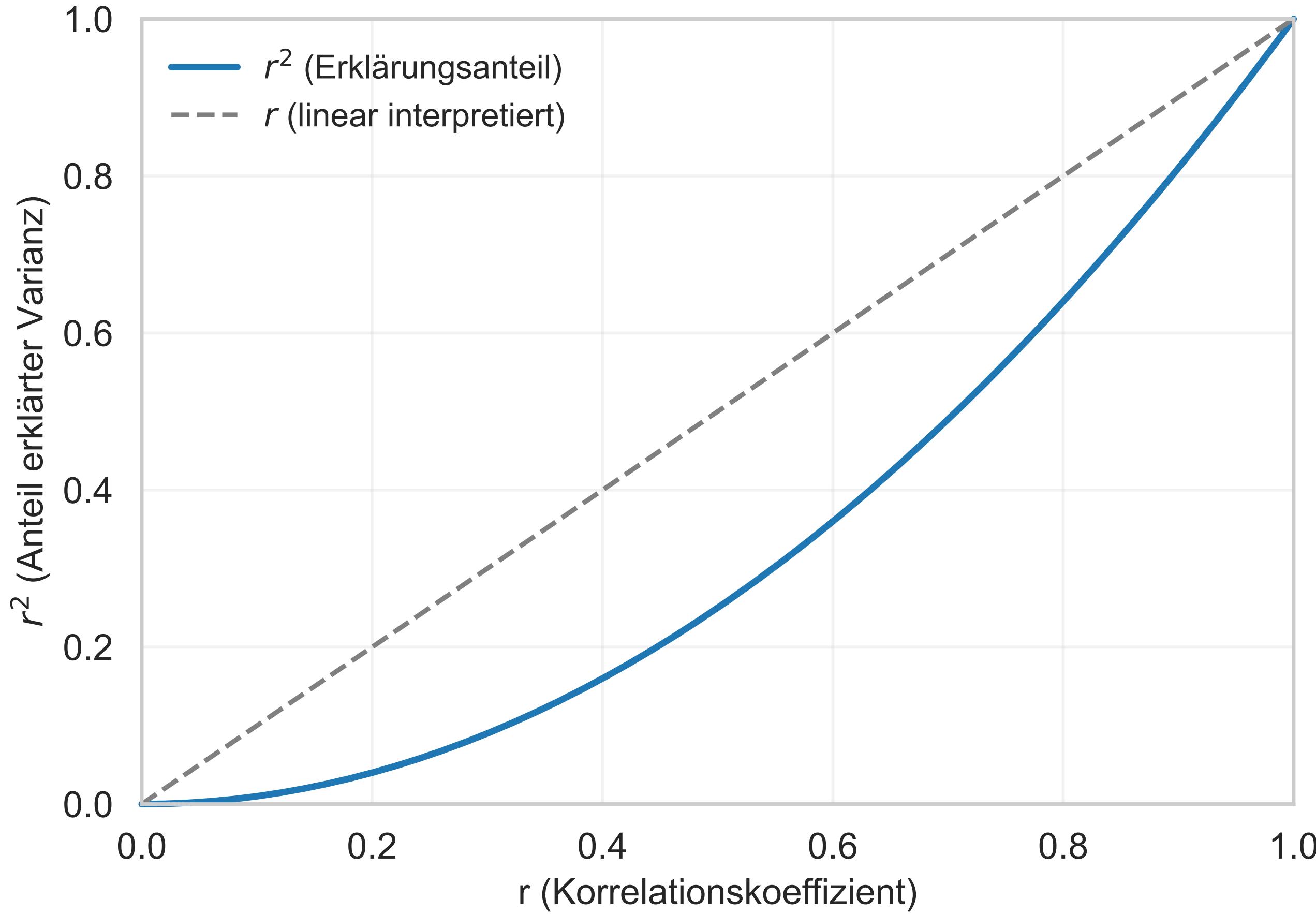
Interpretation von r

FALSCH: $r = 0.8$ ist nicht «80 % Erklärung».

- r misst nur Zusammenhang.
- **RICHTIG:** Anteil erklärter Varianz $= r^2$. (Das Bestimmtheitsmass).
- Beispiel: $r = 0.8 \Rightarrow r^2 = 0.64 \Rightarrow 64\%$ der Varianz in Y wird durch X erklärt.
- Rest = Rauschen, oder andere Einflüsse, die du noch nicht kennst.

Mini-Check: Wie viel Varianz erklärt $r = 0.5$?

Unterschied zwischen r und r^2



Anteil der erklärten Varianz?

Stell dir vor, du willst verstehen, warum **Mathe-Notenpunkte** (Ratioskala) unterschiedlich sind

- Vielleicht hängt das mit der **Lernzeit** zusammen: aber sicher nicht nur.
- Die **Korrelation r** zeigt, **ob** ein Zusammenhang besteht (z. B. mehr Lernen → bessere Note).
- Das **Bestimmtheitsmaß r^2** zeigt, **wie viel der Unterschiede** in den Noten durch Lernzeit **erklärt werden kann**.

Mini-Check: Kann ein hoher r^2 -Wert beweisen, dass mehr Lernen die Note verbessert?

Beispiel

Wenn $r^2 = 0.64$, dann heisst das:

64 % der Unterschiede in den Noten hängen mit der Lernzeit zusammen.

Die restlichen 36 % kommen von anderen Einflüssen (z. B. Schlaf, Motivation, Glück).

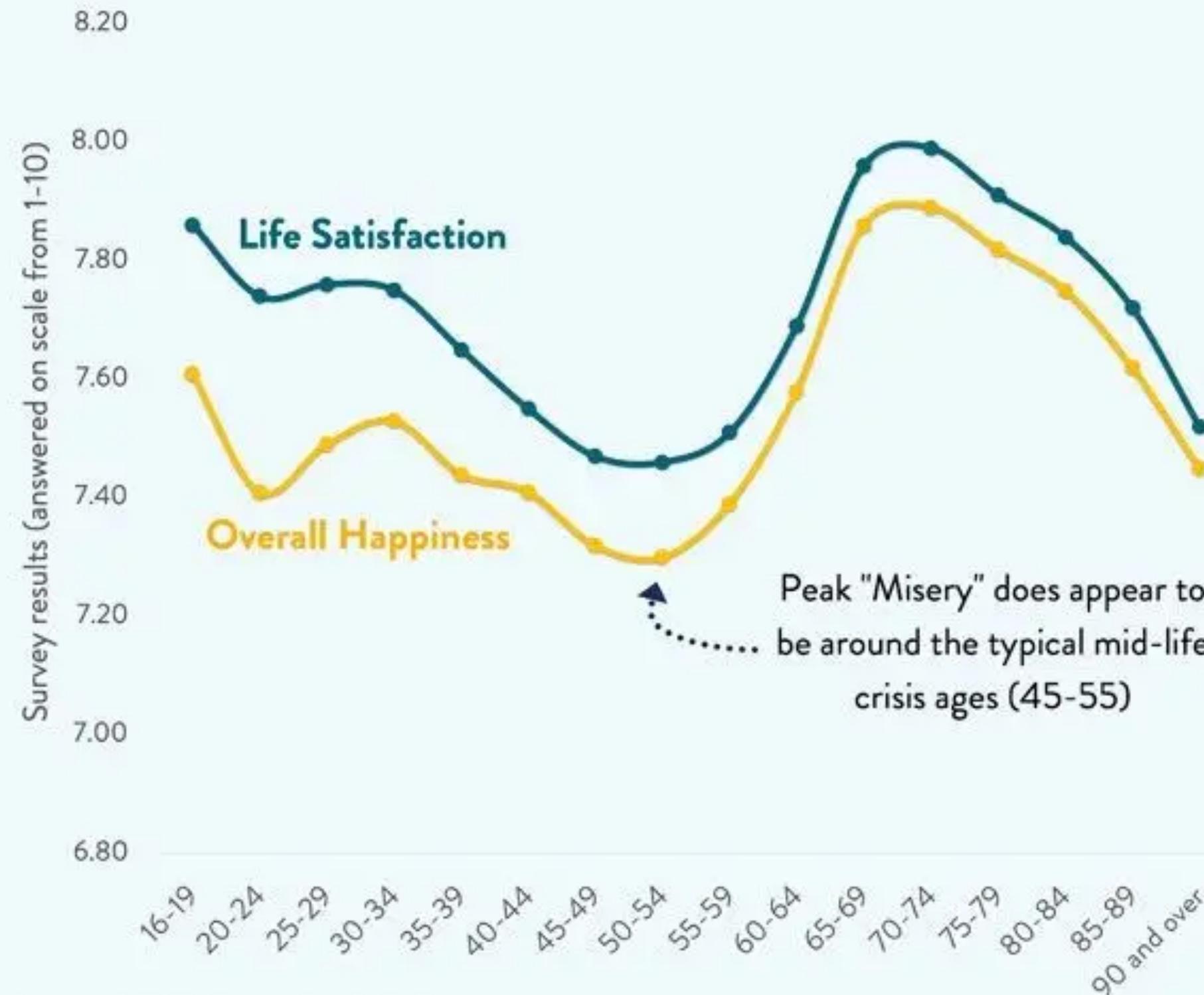
Praxis-Reflexion: Wann nicht Pearson?

Wenn Annahmen verletzt sind, **wechsel das Werkzeug**. Nimm ein Rangmass.

- Nichtlinearer oder nicht normal verteilter Datensatz \Rightarrow Spearman/Kendall.
- Beispiel: Alter und Zufriedenheit (häufig U-Förmig).
- Vorschau: Im nächsten Abschnitt lösen wir dieses Problem.

Mini-Check: Welche Bedingung würde dich zu Spearman führen?

The Happiness Curve



Source: ONS Wellbeing Study (2017)

Sample Size 149,950

Key Takeaways: Kovarianz & Pearson

- **Kovarianz (Grundlage)**: Misst die gemeinsame Abweichung von den Mittelwerten (\bar{x}, \bar{y}). Ist **nicht normiert**.
- **Pearson r (Lineares Mass)**: Ist die standardisierte Kovarianz. Erzeugt einen skalenunabhängigen Wert in $[-1, 1]$.
- **Interpretation r^2** : Der Korrelationskoeffizient r ist **nicht** der Erklärungsanteil. Diesen liefert r^2 (Bestimmtheitsmass).
- **Vorsicht: Aussreisser**: Pearson r ist extrem sensitiv gegenüber Aussreisern und Nicht-Normalität. Ein einziger Punkt kann das Ergebnis verfälschen.
- **Voraussetzung**: Pearson r sollte nur verwendet werden, wenn der Zusammenhang **linear** ist. Nicht-lineare, aber monotone Muster erfordern Rangmasse.

Spearmann & Kendall

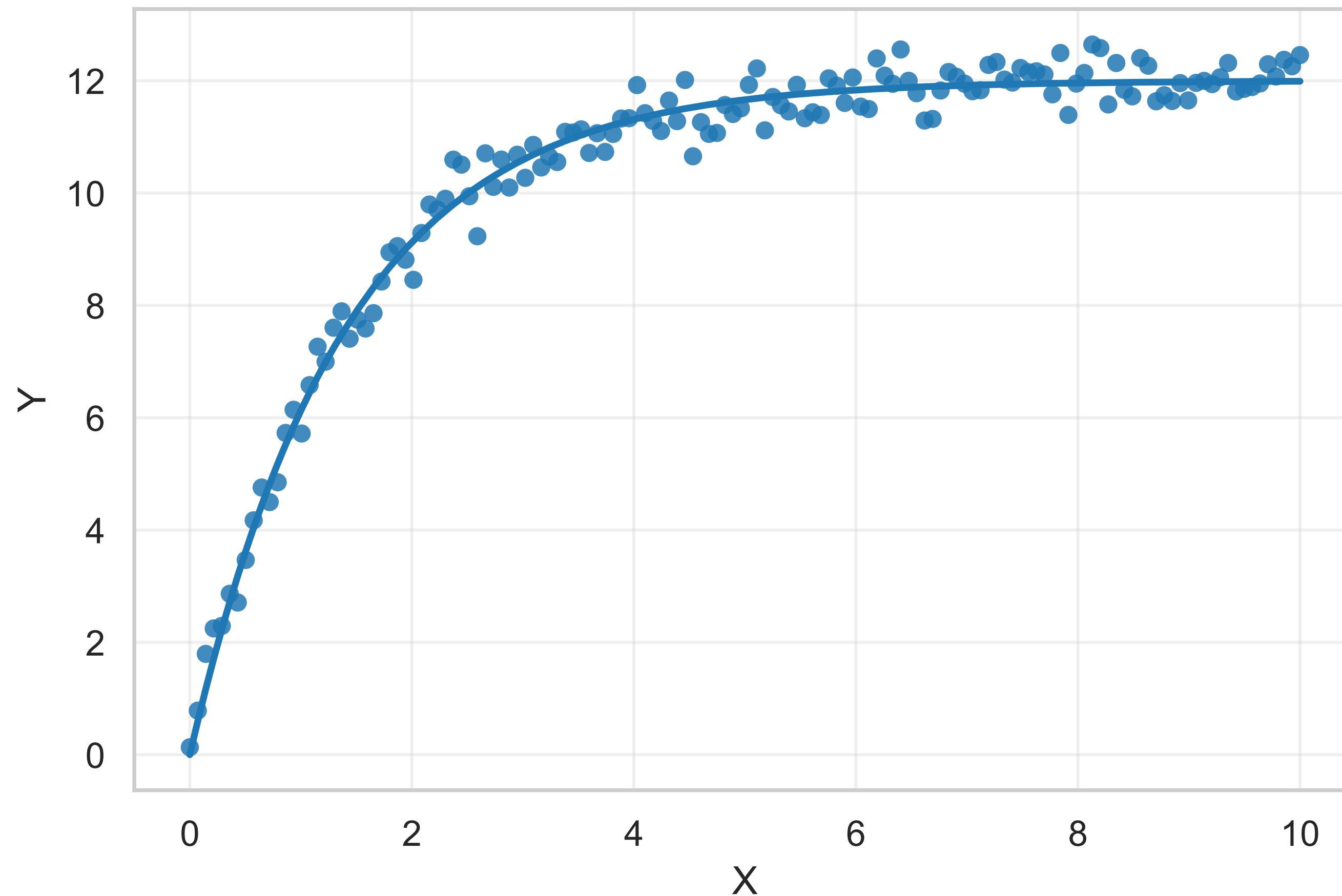
Motivation für Rangkorrelationen

Du brauchst Rangkorrelationen, wenn du **robust** sein musst oder der Zusammenhang nicht linear ist.

- Pearson misst nur **lineare** Trends.
- Bei Verzerrungen oder nicht-metrischen Skalen versagt r .
- Rangkorrelationen (Spearman, Kendall) \Rightarrow basieren auf **relativer Ordnung** (Wer ist Erster, Zweiter, Dritter?).
- Vorteil: Weniger anfällig gegenüber Extremwerten: sie sind nur «Rang 1».

Mini-Check: Warum sind Ränge robuster als Originalwerte?

Monoton & stark nichtlinear: $r=0.78$, $\rho=0.90$



Idee: Ränge statt Werte

Spearman ersetzt deine Rohwerte durch Ränge. Das ist der ganze Trick.

- Jeder Wert \Rightarrow Rangposition im Datensatz.
- Beispiel: Einkommen $[30k, 50k, 70k] \Rightarrow$ Ränge $[1, 2, 3]$.
- Danach rechnest du einfach **Pearson r** auf diesen Rängen.
- Formell: $\rho = \text{corr}(\text{rank}(X), \text{rank}(Y))$.

Mini-Check: Was passiert bei gleichen Werten («Ties»)?

Gleiche Werte («Ties») bei Rängen

Wenn Werte gleich sind, teilen sie sich **denselben Durchschnittsrang**.

Wert	Rang (ohne Ties)	Rang (mit Ties)
10	1	1
20	2	2.5
20	3	2.5
40	4	4

$$(2 + 3)/2 = 2.5$$

Ergebnis: gleiche Werte erhalten denselben Durchschnittsrang: die Reihenfolge bleibt erhalten.

Beispiel: Rangbildung

Rangumwandlung ist einfach nachvollziehbar.

- $X = [2, 4, 6, 8] \Rightarrow$ Ränge $[1, 2, 3, 4]$.
- $Y = [5, 7, 6, 8] \Rightarrow$ Ränge $[1, 3, 2, 4]$.
- Danach Pearson auf Rängen: $\rho \approx 0.8$.
- Zeigt robusten, monotonen Trend.

Mini-Check: Wie ändert sich ρ bei Tausch von $Y = [8, 6, 7, 5]$?

Formel für Spearman ρ

ρ berechnet sich aus Rangdifferenzen.

- Formel (ohne Ties): $\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$
- $d_i = \text{Rang}(x_i) - \text{Rang}(y_i)$.
- Super einfach per Hand berechenbar bei $n < 10$.
- $\rho \approx 1 \Rightarrow$ ähnliche Rangordnung, $\rho \approx -1 \Rightarrow$ umgekehrt.

Mini-Check: Warum ist $\rho = 1$ bei perfekt gleicher Rangordnung?

i	Rang(X)	Rang(Y)	$d_i = X - Y$	d_i^2
1	1	1	0	0
2	2	3	-1	1
3	3	2	1	1
4	4	5	-1	1
5	5	4	1	1

$$\sum d_i^2 = 4 \rightarrow \rho = 1 - (6 \cdot \sum d_i^2) / [n(n^2-1)] = 0.80$$

Bei perfekter Rangordnung (alle $d_i = 0$) wäre $\rho = 1$

Spearman in Python (1)

scipy.stats.spearmanr ist dein Freund.

```
import scipy.stats as st  
x=[10,20,30,40]  
y=[12,22,33,43]  
rho,p=st.spearmanr(x,y)  
print(rho,p)
```

1.0 0.0

Spearman in Python (2)

1.0 0.0

- Ausgabe: $\rho \approx 1, p \approx 0.$
- Vergleich das immer mit `pearsonr`.

Mini-Check: Wann würden ρ und r nahe beieinander liegen?

Vorteile von Spearman

Spearman ist der bessere Allrounder und erkennt mehr Formen.

- Erfasst monotone, **nicht-lineare** Trends.
- Weniger empfindlich gegen **Aussreisser**.
- Keine Normalverteilungs-Annahme nötig.
- Einfache Interpretation («steigt mit steigendem Rang»).

Mini-Check: Was ist der Preis für mehr Robustheit?

Kendall's Tau: Prinzip

Kendall vergleicht Paarordnungen – das ist intuitiv verständlicher.

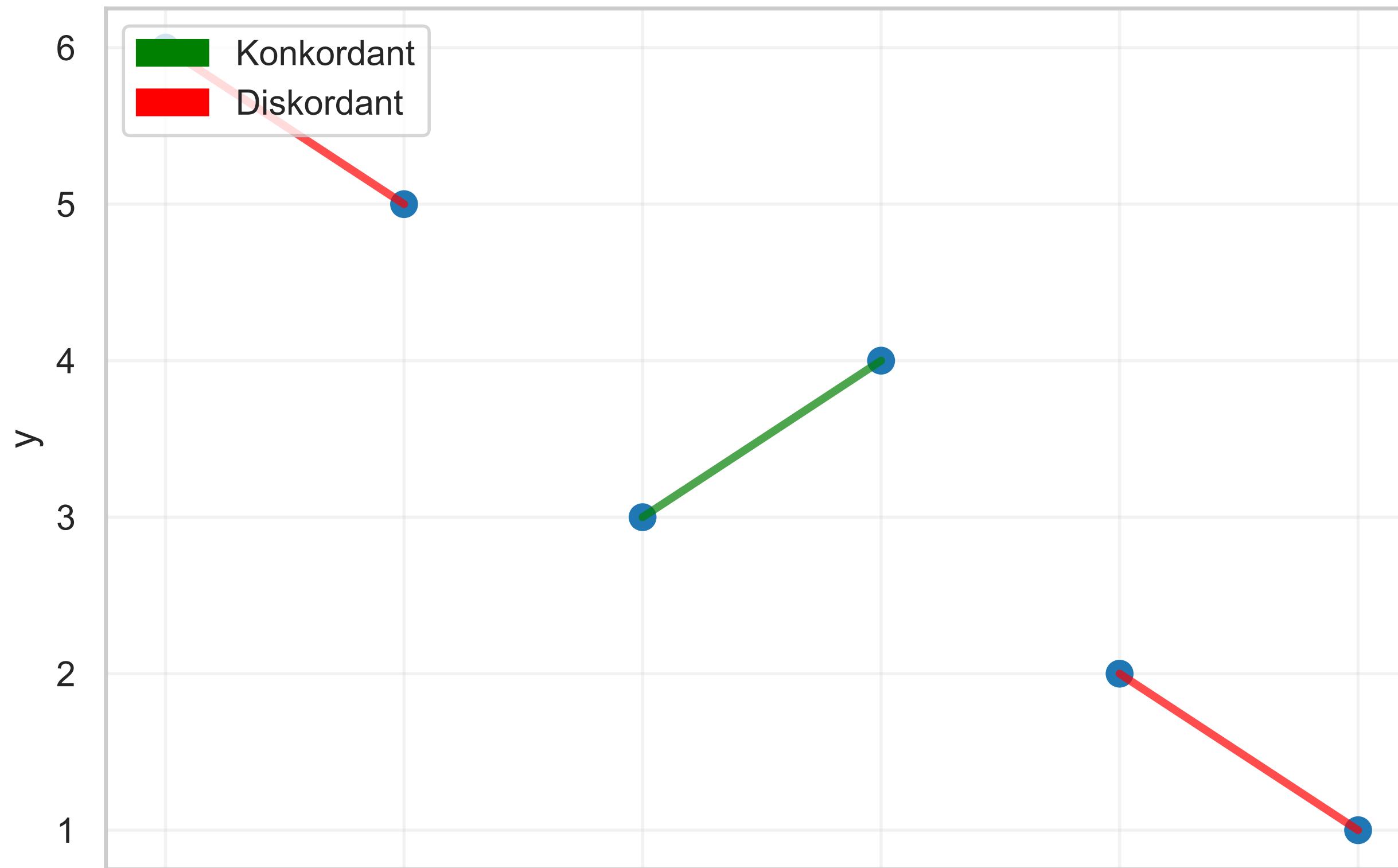
Für jedes Paar (i, j) : prüfe **Konkordanz** (gleiche Ordnung) oder **Diskordanz**.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n - 1)}.$$

$\tau \approx \rho$ bei grossen n , aber **stabiler bei kleinen Stichproben und vielen «Ties»**.

Mini-Check: Was heisst «konkordant» in diesem Kontext?

Kendall's Tau – Konkordante und Diskordante Paare



$$\begin{aligned}\tau &= \frac{n_c - n_d}{\frac{1}{2}n(n-1)} & 2 \\ &= \frac{1 - 14}{15} = -0.87\end{aligned}$$

Vergleich Spearman vs. Kendall

Beide messen Monotonie, aber unterschiedlich sensitiv.

- Typisch: $\rho \approx 0.9$, $\tau \approx 0.7$. (τ ist immer gedämpfter.)
- Kendall $\tau \Rightarrow$ besser bei vielen **Ties** (gleichen Rängen).
- Spearman $\rho \Rightarrow$ schneller bei grossen n .
- Beide sind robust gegen Aussreisser.

Mini-Check: Wann würdest du τ statt ρ verwenden?

Beispiel: Kendall in Python

Auch Kendall ist schnell im Code erledigt.

```
st.kendalltau(x,y)
```

- liefert τ und p-Wert.
- Vergleich mit Spearman aus vorheriger Folie.
- Kommentar: τ liefert ähnliche Erkenntnisse, ist aber etwas gedämpfter im Wert.

Mini-Check: Wie unterscheidet sich τ numerisch von ρ bei Rauschen?

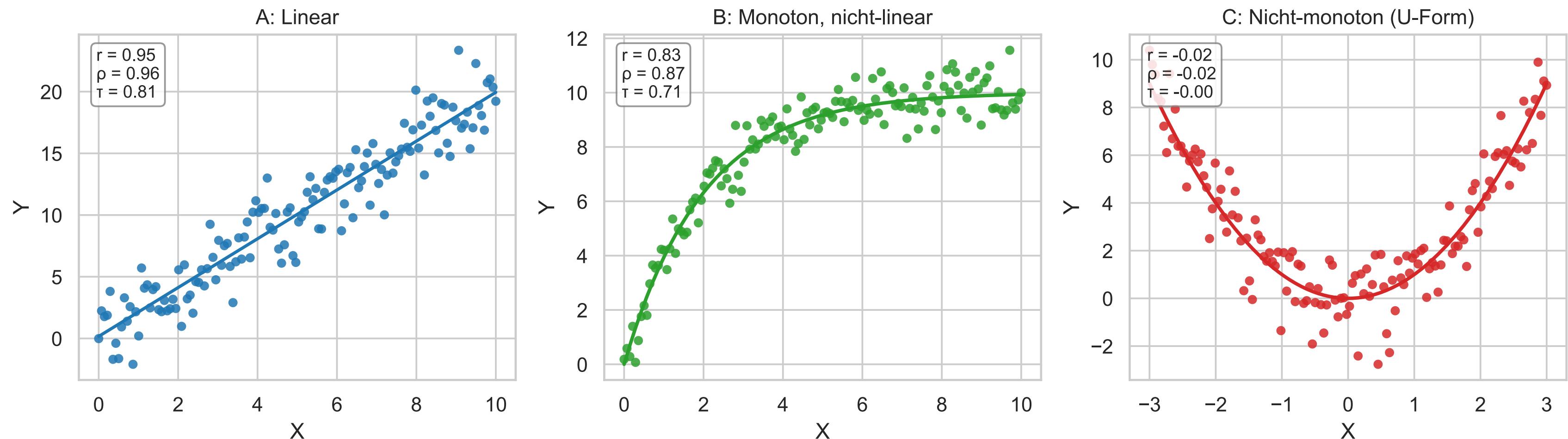
Übungsfrage: Vergleich drei Korrelationen

Die richtige Wahl des Masses entscheidet über die richtige Interpretation.

- Stell dir drei Datensätze (A, B, C) vor: linear, monoton, nicht-monoton.
- Schätze Pearson r , Spearman ρ , Kendall τ .
- Kurze Gruppen-Diskussion (2 min).

Mini-Check: Welcher Datensatz hat höchste ρ , welcher höchste r ?

Vergleich drei Korrelationen: Pearson r , Spearman ρ , Kendall τ



Key Takeaways: Rangkorrelationen

- Wann? Rangkorrelationen, wenn Pearson versagt: also bei nichtlinearer Form, Schiefe oder Aussreissern.
 - Spearman (ρ): Korreliert Ränge statt Werte → misst monotone Trends.
 - Kendall (τ): Vergleicht Paarordnungen → stabil bei kleinen n & vielen Gleichständen.
 - Vorteil: Robust gegen Extremwerte, keine Normalverteilungsannahme nötig.
 - Praxis: Ideal für ordinale Daten oder gleichgerichtet gekrümmte Verläufe.
- 👉 Plotte die Daten → Always plot your data!
- 👉 Wenn es gekrümmt oder robust nötig ist → nimm Spearman.

Visualisierung von Korrelationen

Warum Visualisierung entscheidend ist

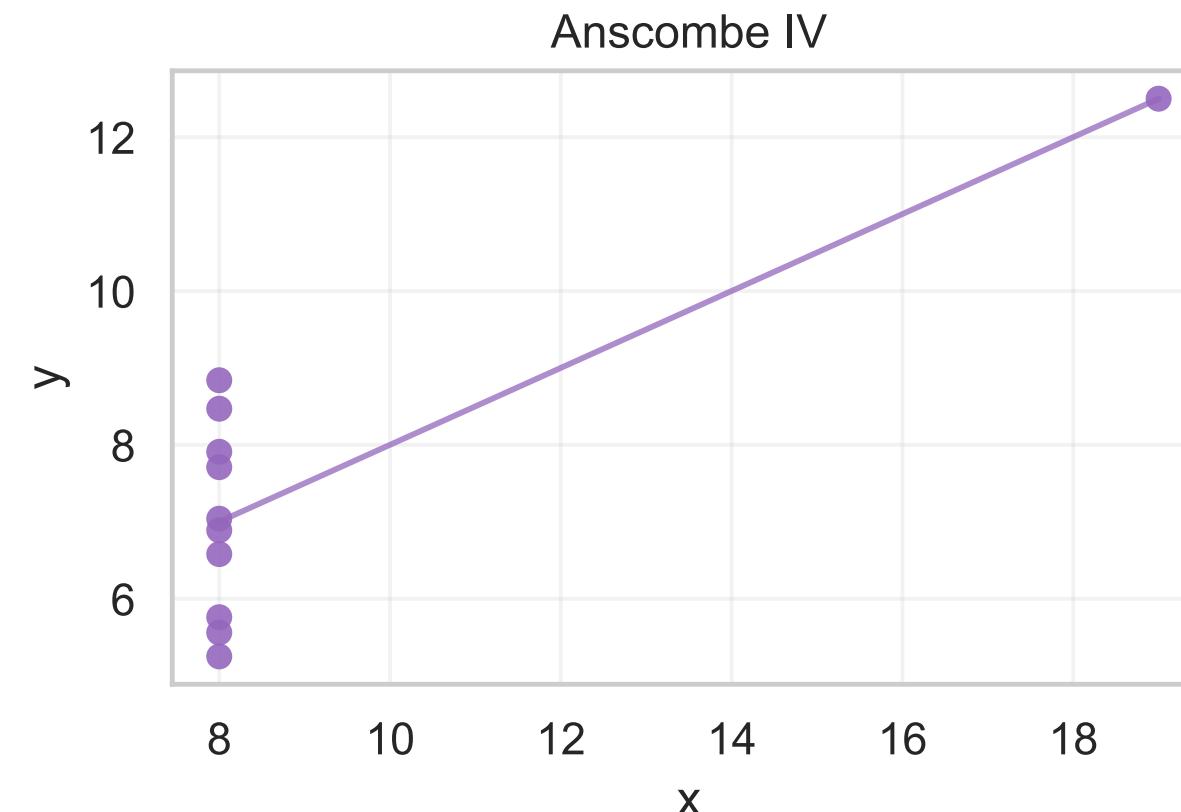
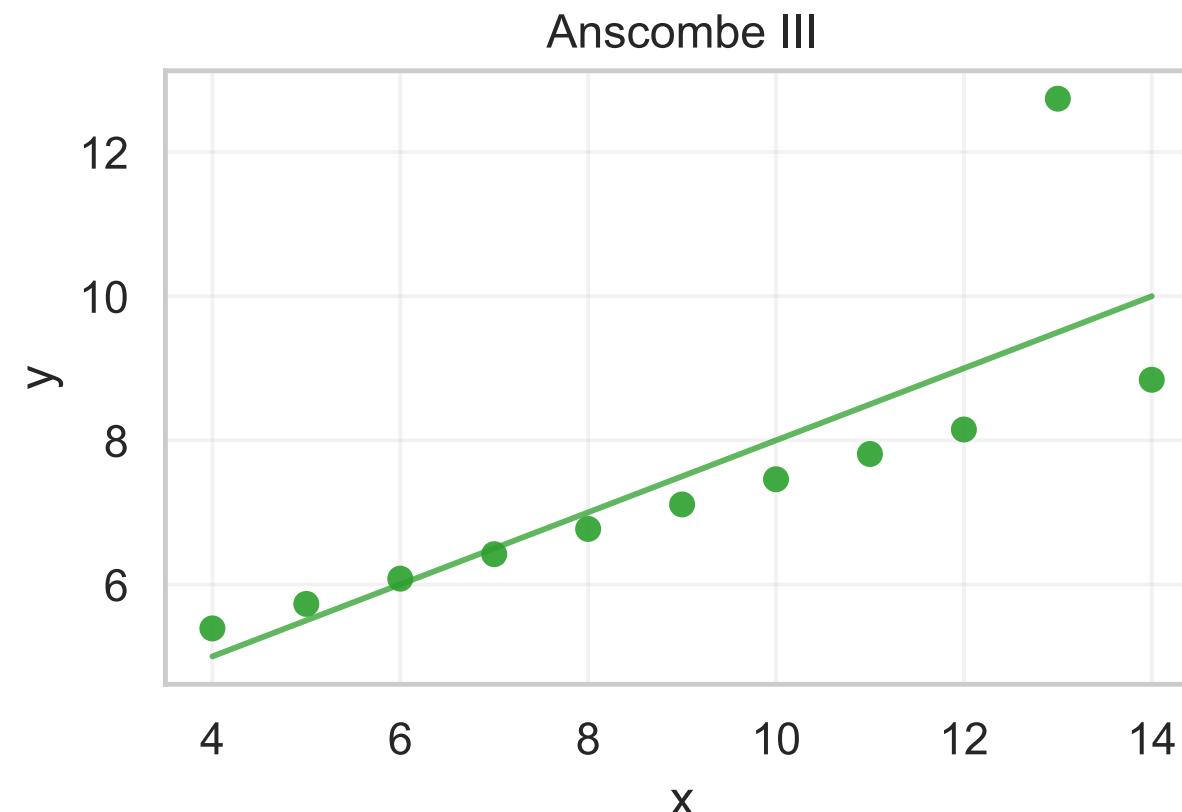
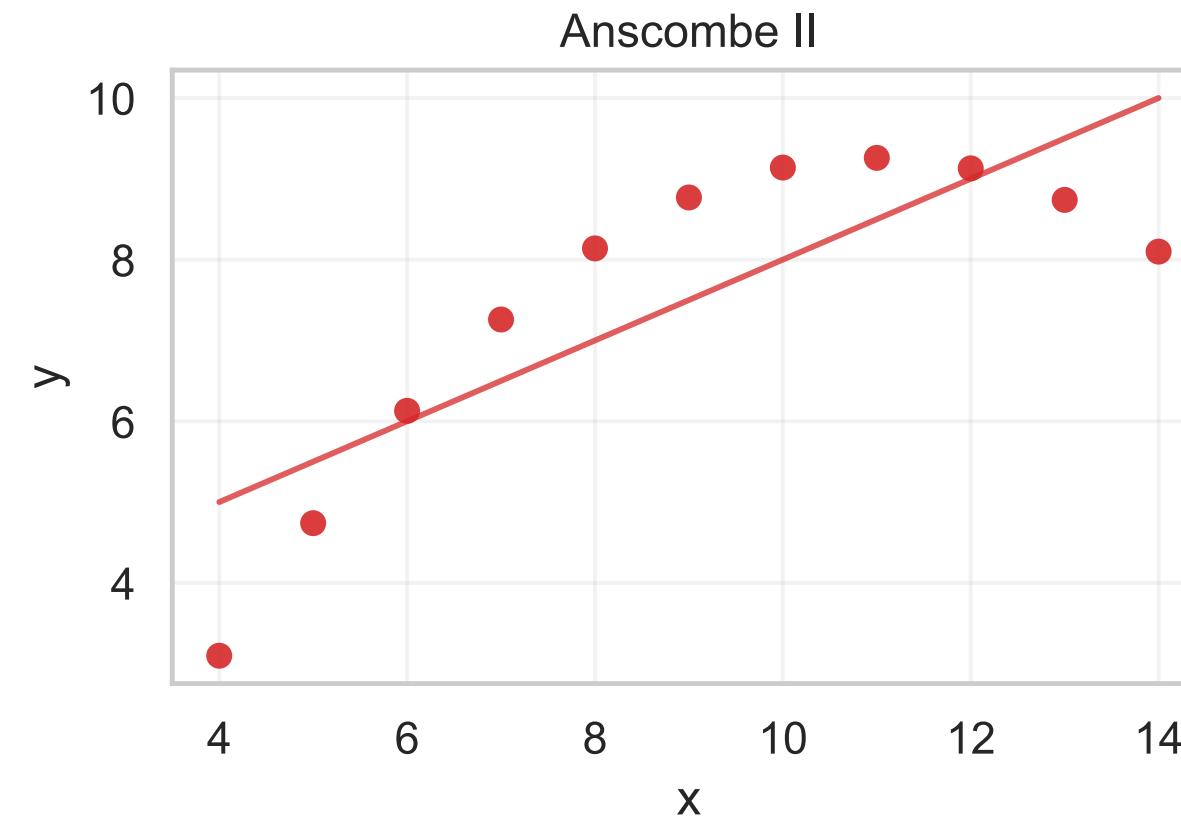
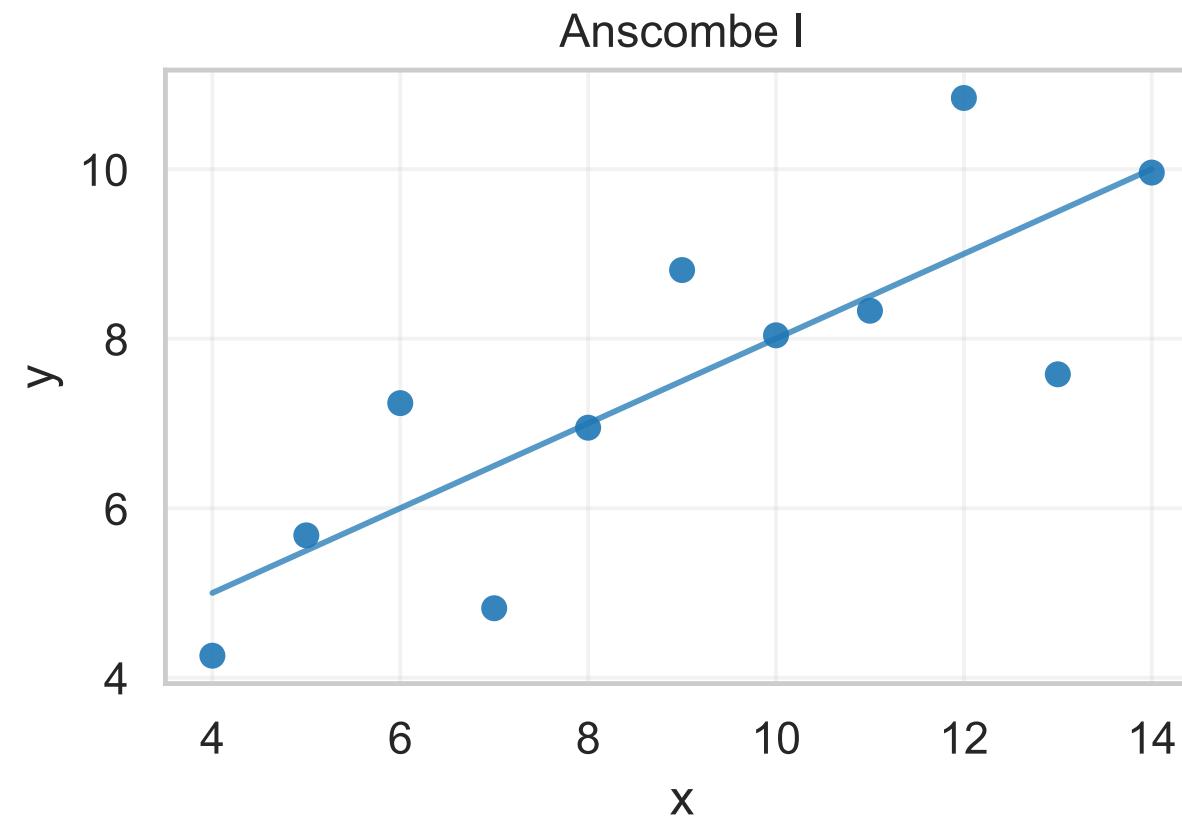
Ein Plot zeigt Zusammenhänge, die Zahlen **verbergen**.

- Der Schock: Gleiches $r \Rightarrow$ sehr unterschiedliche Datenstrukturen (Anscombe-Quartett).
- **Deine Pflicht:** Aussreißer, Cluster, Nichtlinearitäten werden nur **visuell** sichtbar.
- Korrelation immer mit Plot prüfen: das ist der **Pflichtschritt** in der EDA (Exploratory Data Analysis).

«Trust, but plot.»

Mini-Check: Was würde ein Scatterplot zeigen, was r **nicht** zeigt?

Anscombe's Quartett – gleiche Statistik, unterschiedliche Struktur



Scatterplot: Die Grundform

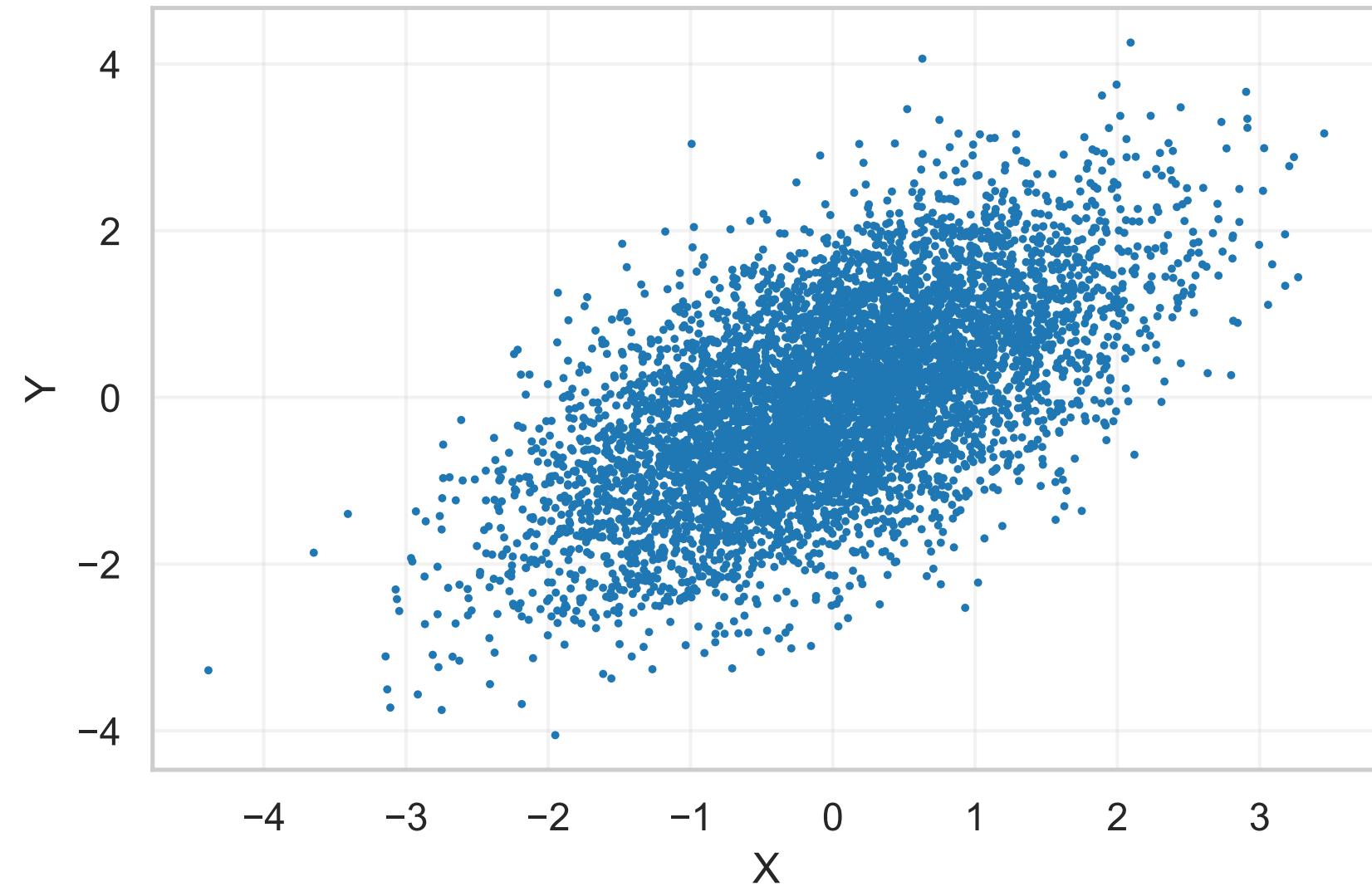
Scatterplots sind das **Standard-Werkzeug** für bivariate Beziehungen.

- Achse X vs. Achse Y : Jeder Punkt ist eine Beobachtung.
- Dichte zeigt den Trend, die Form zeigt die Art des Zusammenhangs (linear vs. kurvig).
- Bei grossem n : **Overplotting** beachten \Rightarrow nutze Transparenz (alpha) oder Hexbins.

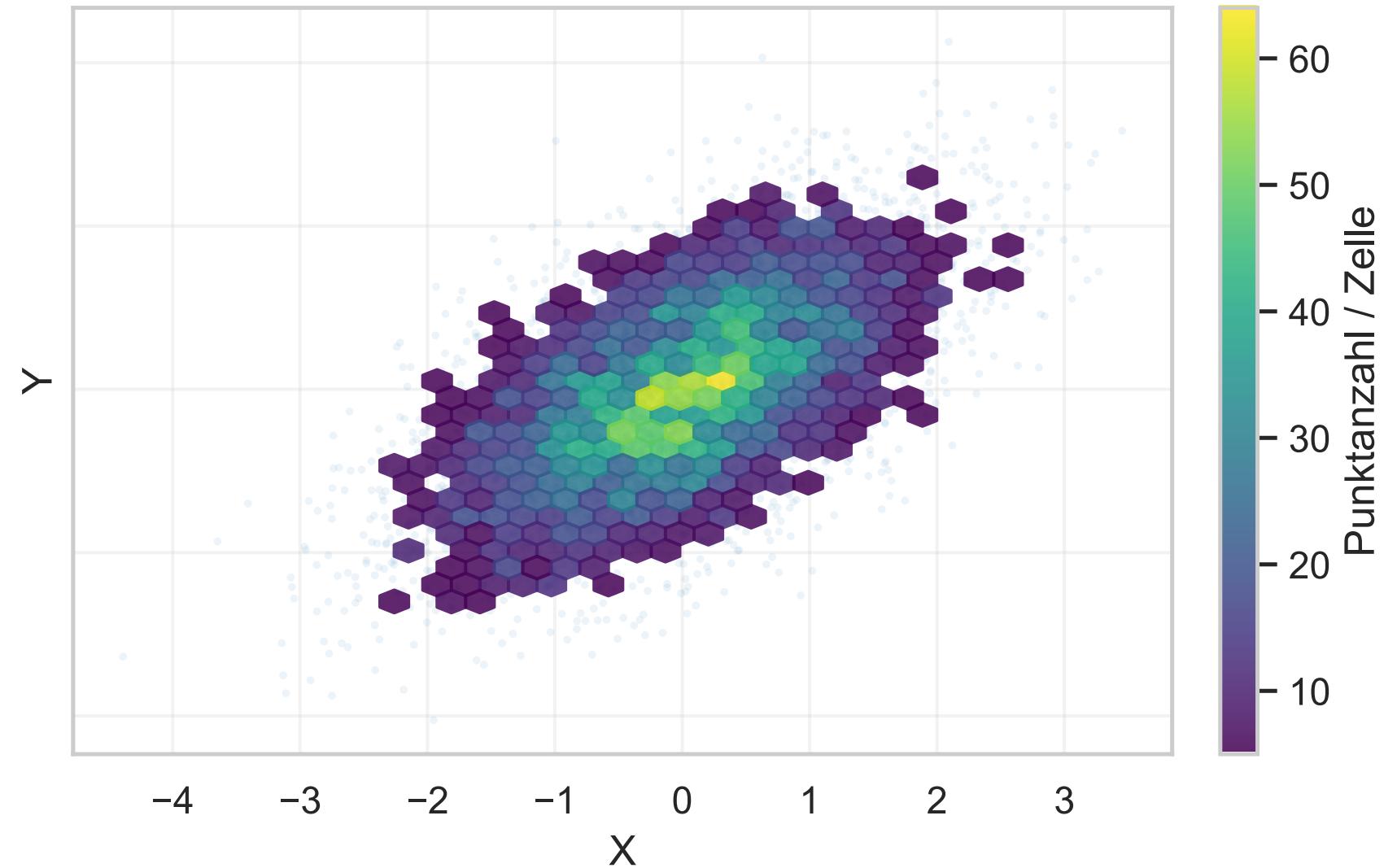
Mini-Check: Wie kannst du Overplotting reduzieren?

Scatterplot: Struktur sichtbar machen – Overplotting vermeiden

Overplotting: viele Punkte, keine Transparenz



Gegenmittel: Transparenz + Hexbins

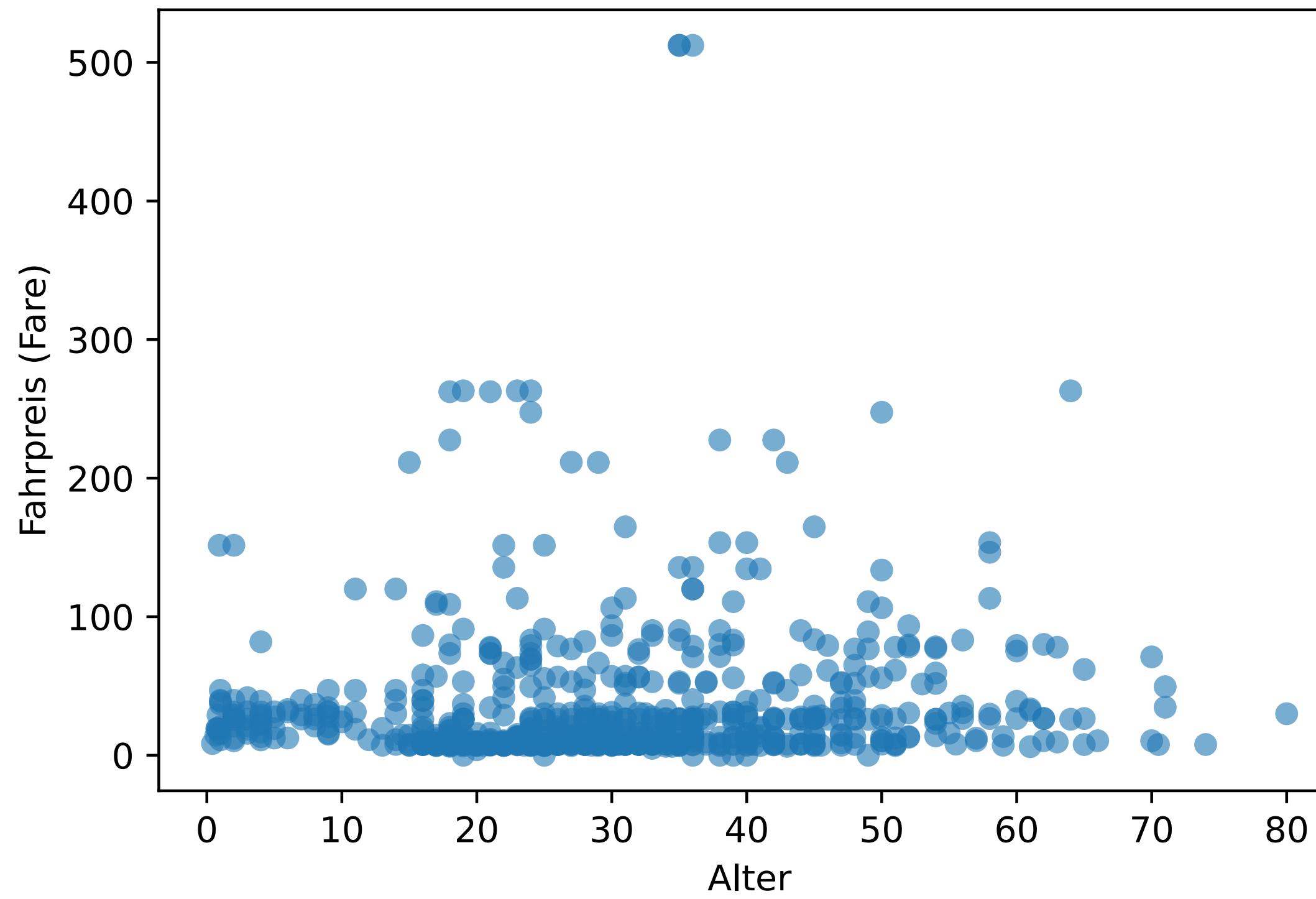


Python: Scatterplot in Seaborn

Seaborn bietet dir eine komfortable Funktion für Streudiagramme.

```
import seaborn as sns
titanic = sns.load_dataset("titanic")
# Alpha reduziert Overplotting
sns.scatterplot(x="age", y="fare", data=titanic, alpha=0.6)
```

Scatterplot – Titanic: Alter vs. Fahrpreis



Python: Scatterplot in Seaborn

```
import seaborn as sns  
titanic = sns.load_dataset("titanic")  
# Alpha reduziert Overplotting  
sns.scatterplot(x="age", y="fare", data=titanic, alpha=0.6)
```

- Optional: hue für Farbkodierung nach einer dritten Variable.
- Interpretation: Du siehst einen leicht positiven Trend, aber mit vielen, vielen Aussreissern.

Mini-Check: Wie würdest du den Plot lesen, wenn $r = 0.2$ ist?

Regressionslinie im Scatterplot

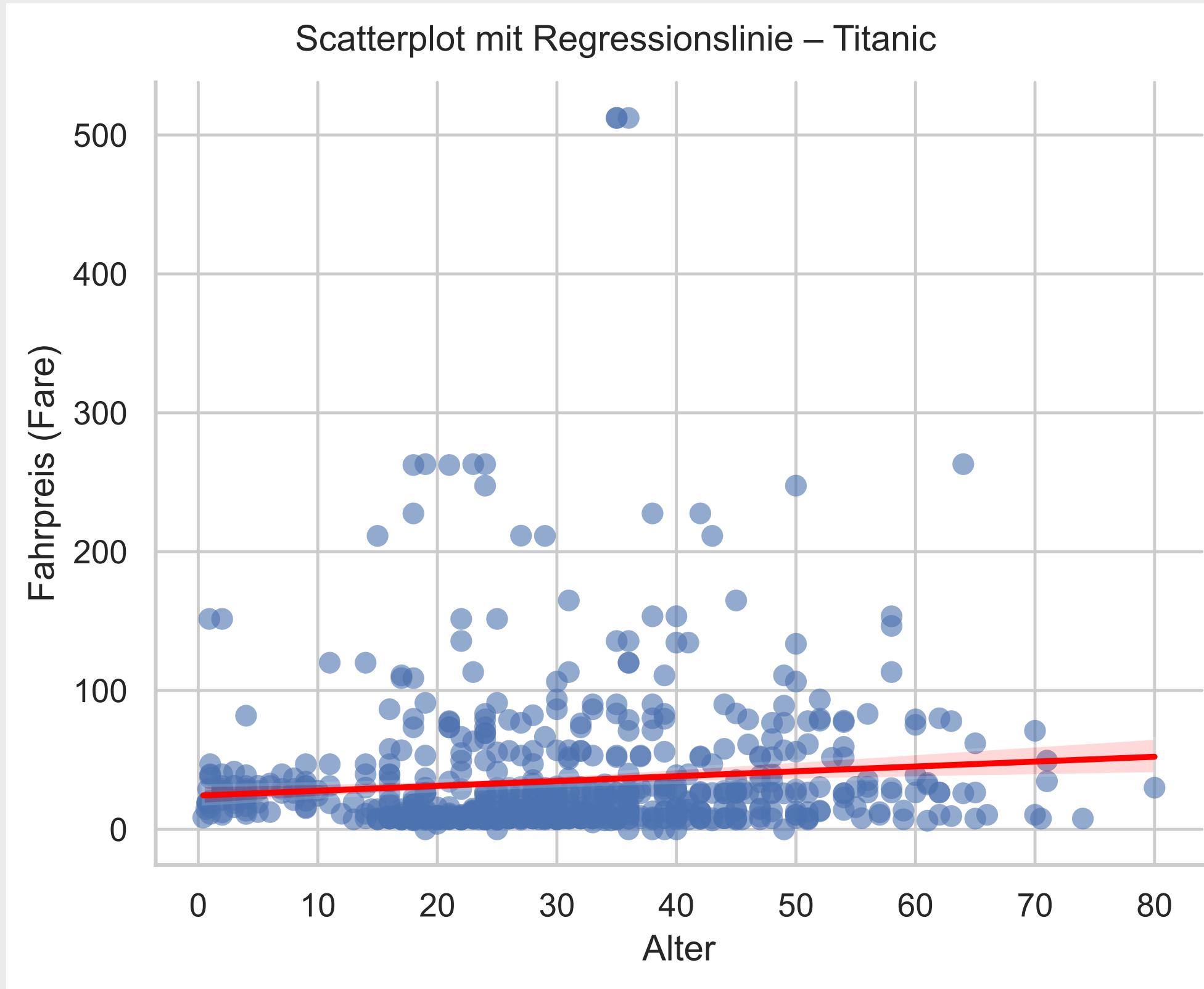
Ein Fit verdeutlicht den Trend, aber impliziert **keine Kausalität**.

```
sns.lmplot(x="age", y="fare", data=titanic)
```

- Die lineare Regression zeigt dir die **optimale Gerade** zur Orientierung.
- Achtung: Es ist **nicht** «Vorhersage», sondern nur Zusammenhangsanzeige.
- Das Konfidenzintervall (der Schatten) zeigt dir die **Unsicherheit** des Trends.

Mini-Check: Wann würde eine deutliche Krümmung auf Nichtlinearität hinweisen?

Scatterplot mit Regressionslinie – Titanic



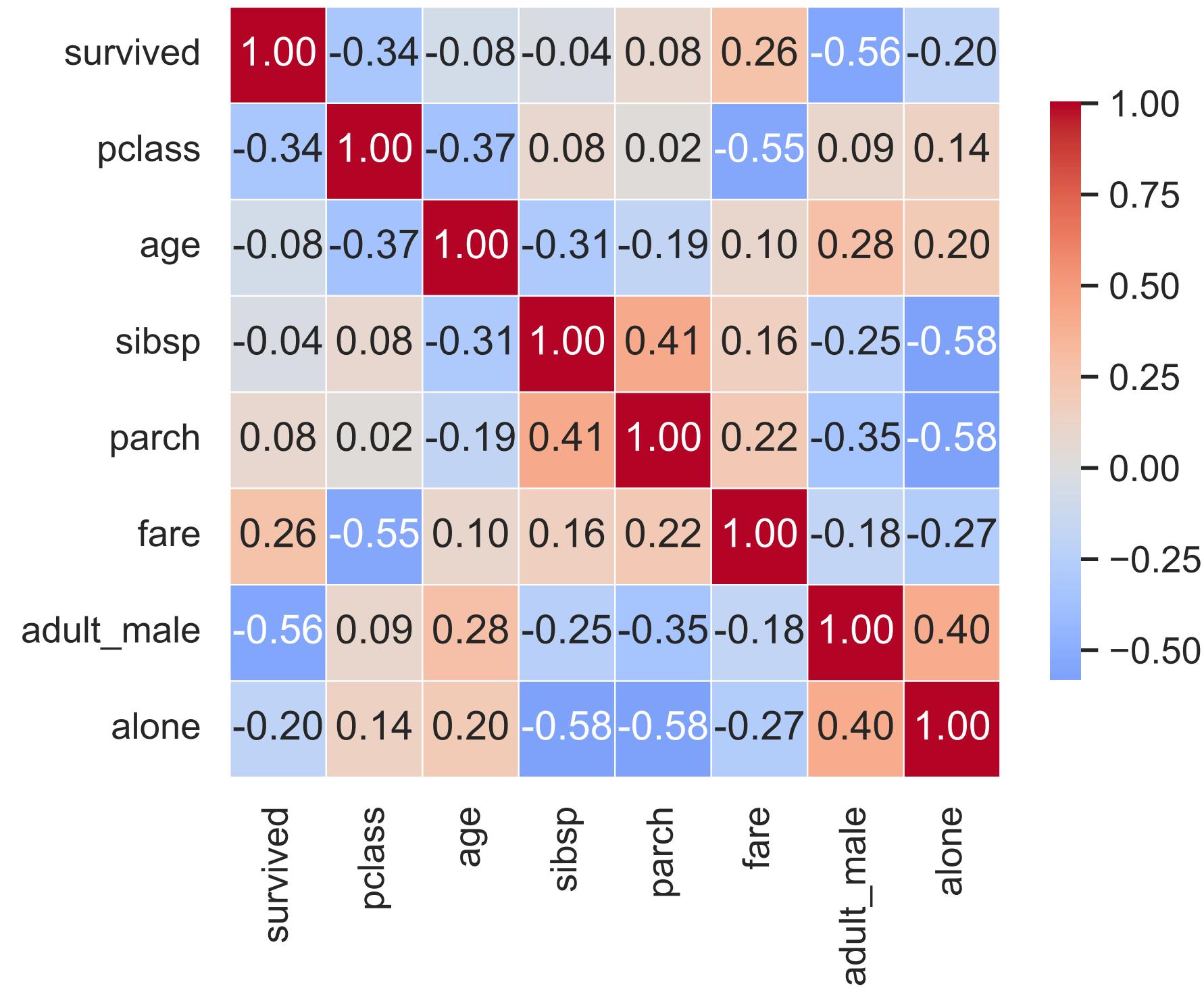
Heatmap: Korrelationsmatrix

Heatmaps zeigen dir **viele Korrelationen** auf einen Blick.

- Farben kodieren Werte (-1 Blau $\Rightarrow +1$ Rot).
- Die Diagonale ist immer 1 (r mit sich selbst).
- Einsatz: Die **erste Strukturprüfung** in der EDA.

```
import seaborn as sns
titanic = sns.load_dataset("titanic")
corr = titanic.corr(numeric_only=True)
sns.heatmap(corr, annot=True, cmap="coolwarm", fmt=".2f", center=0)
```

Korrelationsmatrix – Titanic-Datensatz



Interpretation von Heatmaps

Verlass dich nicht nur auf die Farben: **Kontext entscheidet.**

- Korrelationsstärke musst du immer im Kontext der Domäne sehen.
- Bsp: fare & pclass ($\rho \approx -0.55$) \Rightarrow starker negativer Zusammenhang: höhere Klasse, höherer Preis.
- Versus age & fare ($\rho \approx 0.10$) \Rightarrow praktisch kein linearer Zusammenhang.
- **Vorsicht:** Die Farbskala kann durch Skalierung (z.B. nur $r \in [0.1, 0.4]$) Täuschung verursachen.

Mini-Check: Warum nicht nur auf Farbe verlassen?

Pairplot: Alles auf einen Blick

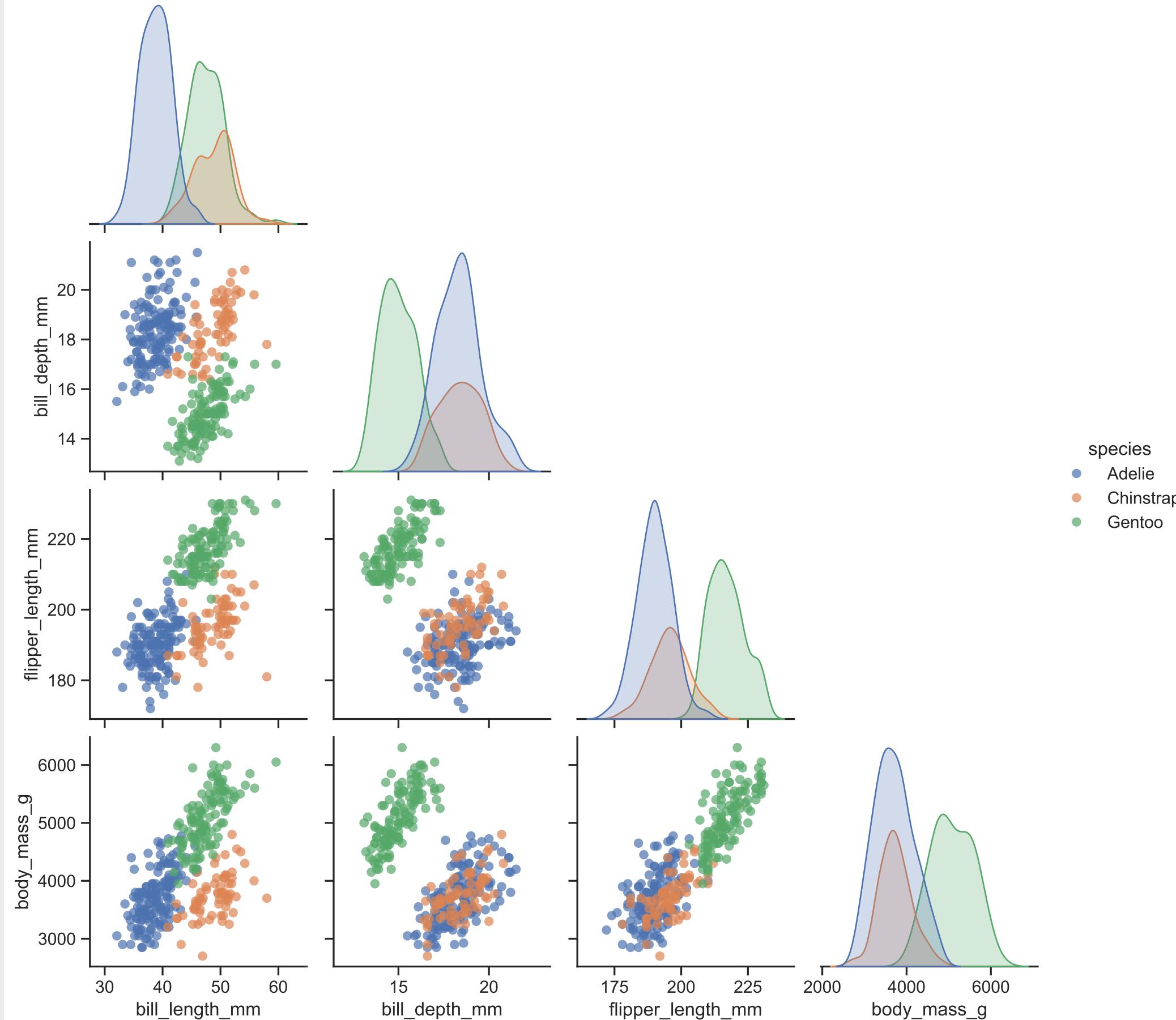
Pairplots zeigen **Scatterplots aller Variablenpaare** in einer einzigen, kompakten Matrix.

```
sns.pairplot(penguins, corner=True)
```

- Auf der Diagonalen: Histogramme oder Dichteplots (Univariate Verteilung).
- Off-Diagonal: Bivariate Beziehungen.
- Super, um Cluster, Nichtlinearitäten und Aussreisser schnell zu erkennen.

Mini-Check: Was sagen dichte elliptische Wolken aus?

Penguins – Paarweise Zusammenhänge der Körpermerkmale



Erweiterte Visuals: Regplot & Jointplot

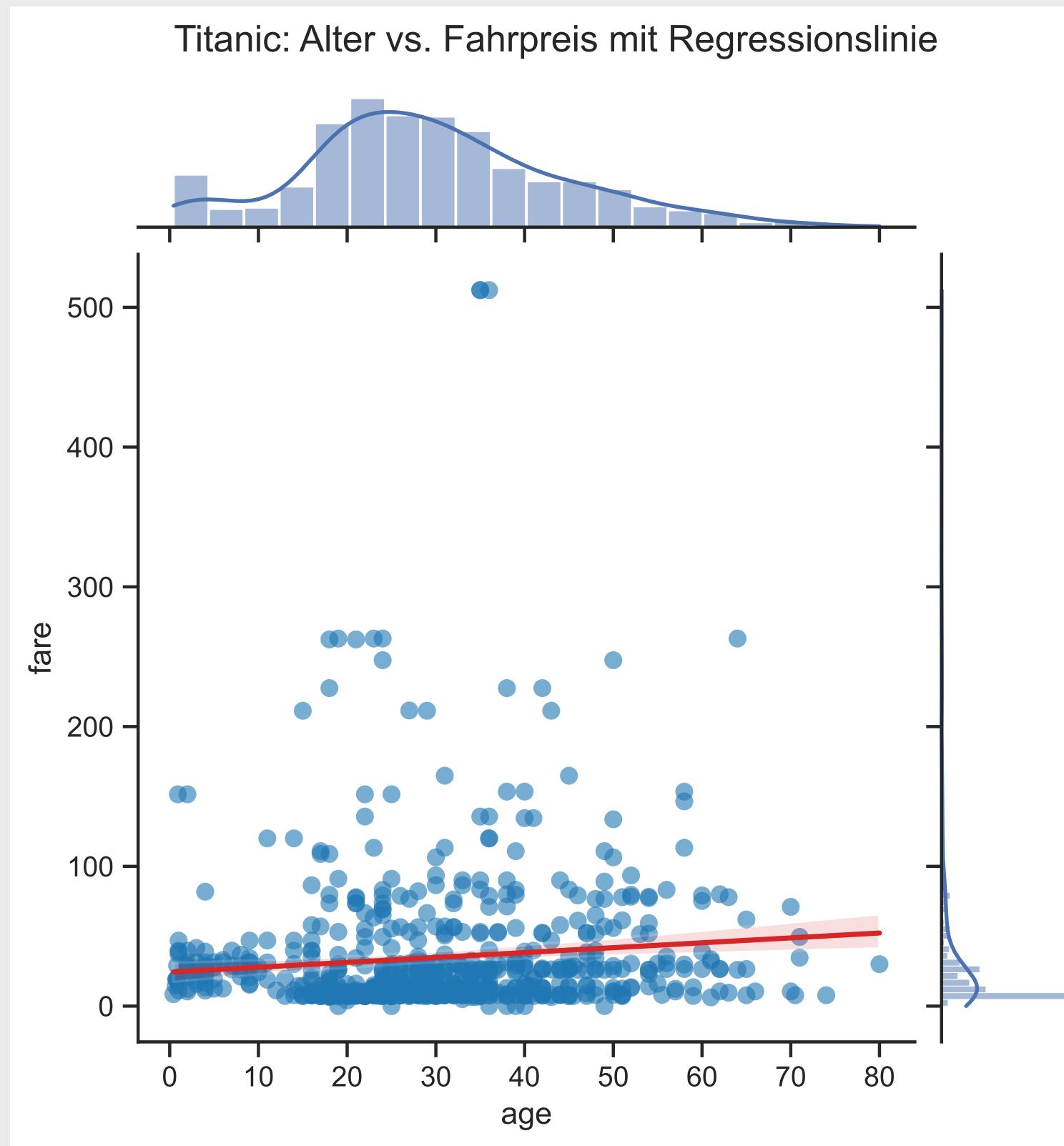
Jointplots kombinieren Scatter + Histogramme + Fit in einem Bild. Effizienz pur.

```
sns.jointplot(x="age", y="fare", data=titanic, kind="reg")
```

- Zentrierte Darstellung der bivariaten Beziehung mit den Randverteilungen (Histogramme) am Rand.
- Praktisch für die Tiefenexploration von Zwei-Variablen-Beziehungen.

Mini-Check: Welcher Plot zeigt Randverteilungen automatisch mit?

Titanic: Alter vs. Fahrpreis mit Regressionslinie



Praxisübung: Korrelation im Datensatz

Visualisierung schärft deine Interpretationskompetenz am besten.

- Aufgabe: Berechne und plotte Korrelationen für das penguins Dataset.
- Identifiziere 3 stärkste Paare (z.B. mit `df.corr().abs().nlargest(3)`).
- Kurz Peer-Austausch (2 min).

Mini-Check: Welche Variablen sind offensichtlich korreliert (z.B. Körpermasse und Schnabellänge)?

Visualisierung: Best Practices

Gute Plots erzählen eine **klare Geschichte**.

- Achsen beschriften und Skalen prüfen (starten die Achsen bei Null?).
- Farben klar und bedeutsam wählen (nicht einfach nur bunt).
- Immer Kontext und Einheiten angeben.
- Das Ziel ist nicht «schön», sondern **«verständlich»** und **«wahrhaftig»**.

Mini-Check: Nenne einen Fehler, den du bei Korrelationsplots vermeiden würdest.

Key Takeaways: Visualisierung (EDA)

- Plotting ist Pflicht
 - Wähle das Werkzeug:
 - Scatterplot: Zeigt die bivariaten Beziehungen am besten (Form, Ausreisser).
 - Heatmap: Zeigt die Korrelationen vieler Variablenpaare auf einen Blick (EDA, Feature-Selektion).
 - Pairplot: Kombination aus Scatterplots und Histogrammen für den besten Gesamtüberblick.
 - Trendlinie prüfen: Die lmplot-Regressionslinie visualisiert den linearen Trend deines Pearson r -Werts, beweist aber keine Kausalität.
 - Best Practice: Achte auf Overplotting (nutze alpha) und darauf, dass deine Plots verständlich und kontextualisiert sind, nicht nur «schön».
- 👉 Sei kritisch: Vertraue keinen Zahlen, die du nicht selbst geplottet hast.

Praxisfallen & Simpson-Paradox

Korrelation \neq Kausalität

Korrelation beschreibt Muster, nicht Ursachen.

- Auch zufällige oder durch Dritte erzeugte Beziehungen ergeben hohe r -Werte.
- Beispiel-Klassiker: Speiseeisverkauf $\uparrow \leftrightarrow$ Ertrinken \uparrow (Z = Temperatur).
- **Deine Regel:** Nur experimentelle Kontrolle erlaubt kausale Schlüsse.
- Die Gefahr: «post hoc \Rightarrow propter hoc» (Nachher \Rightarrow Deswegen).

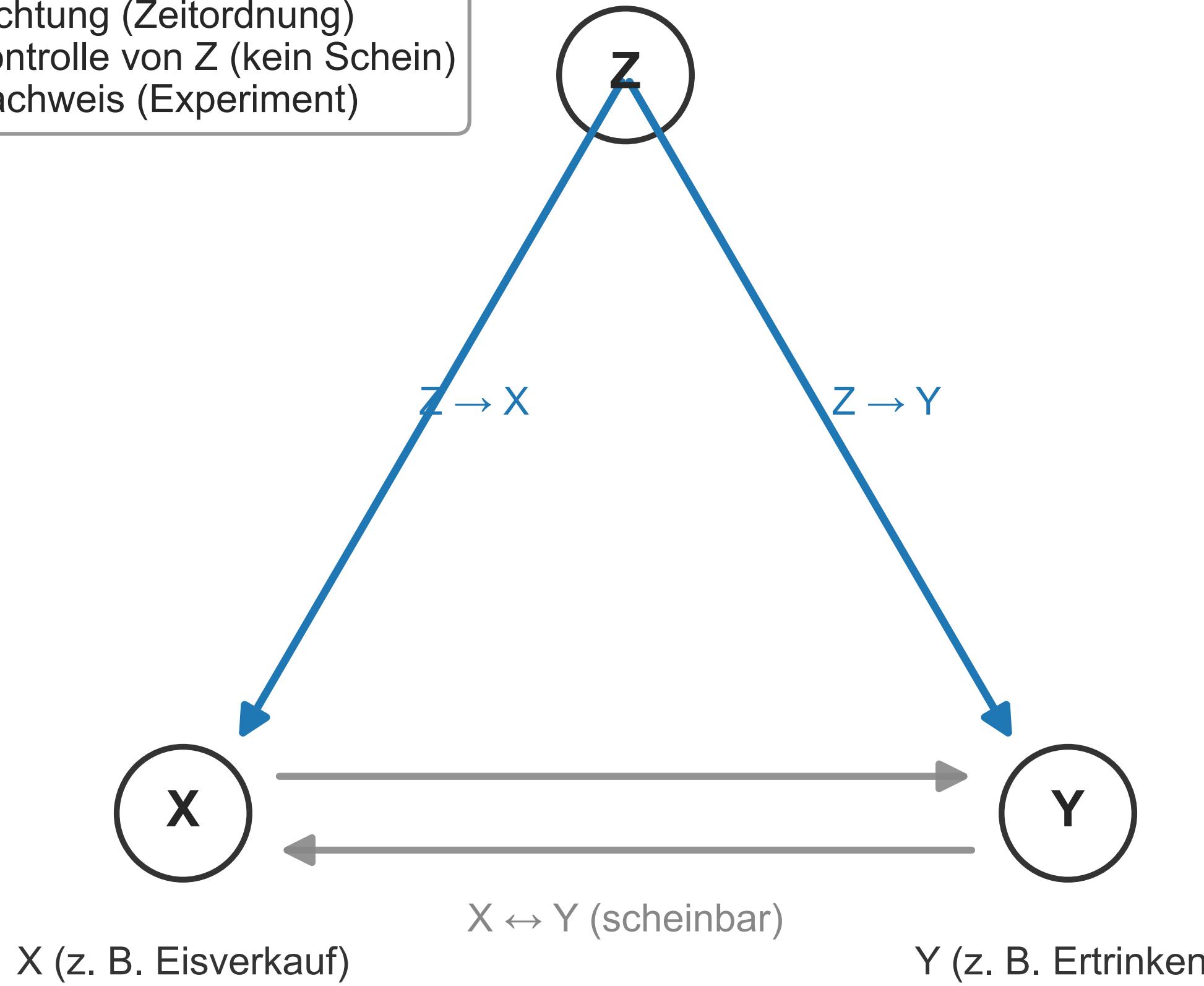
Mini-Check: Was braucht es, um Kausalität nachzuweisen?

Korrelation \neq Kausalität – Konfundierung durch Z

Beispiel: Temperatur

Kausalität braucht:

- Richtung (Zeitordnung)
- Kontrolle von Z (kein Schein)
- Nachweis (Experiment)



Scheinkorrelationen

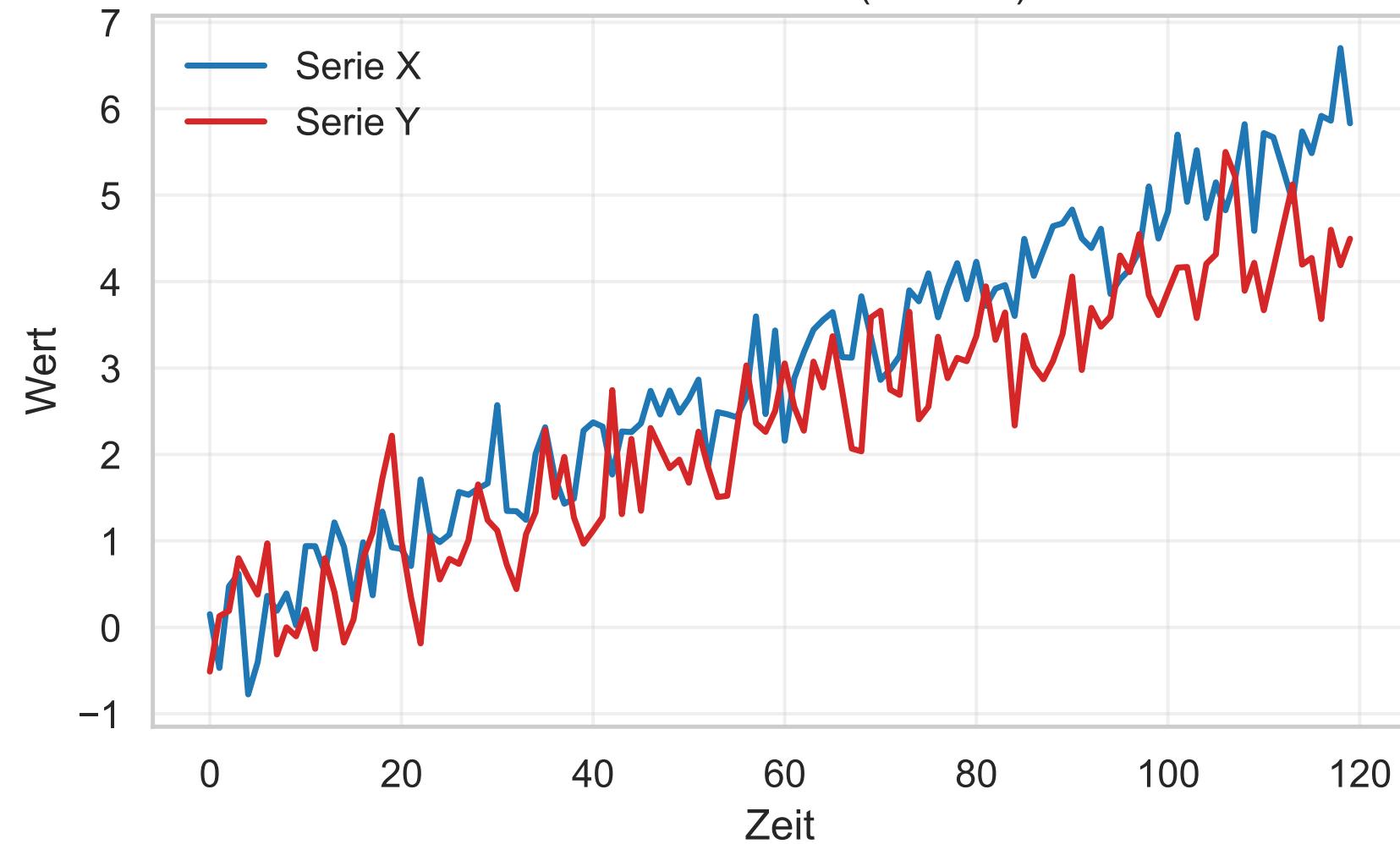
Hohe r -Werte können ohne kausalen oder mechanistischen Zusammenhang entstehen. Das ist Zufall.

- Beispiel: Piratenzahl vs. globale Temperatur (beide sinken über die Zeit).
- Beeinflussung durch **Zeitreihen** oder **gemeinsame Trends** (z.B. GDP-Wachstum treibt alles).
- Warnzeichen: Beide Variablen steigen oder fallen gleichzeitig, aber es gibt keinen **direkten Mechanismus**.

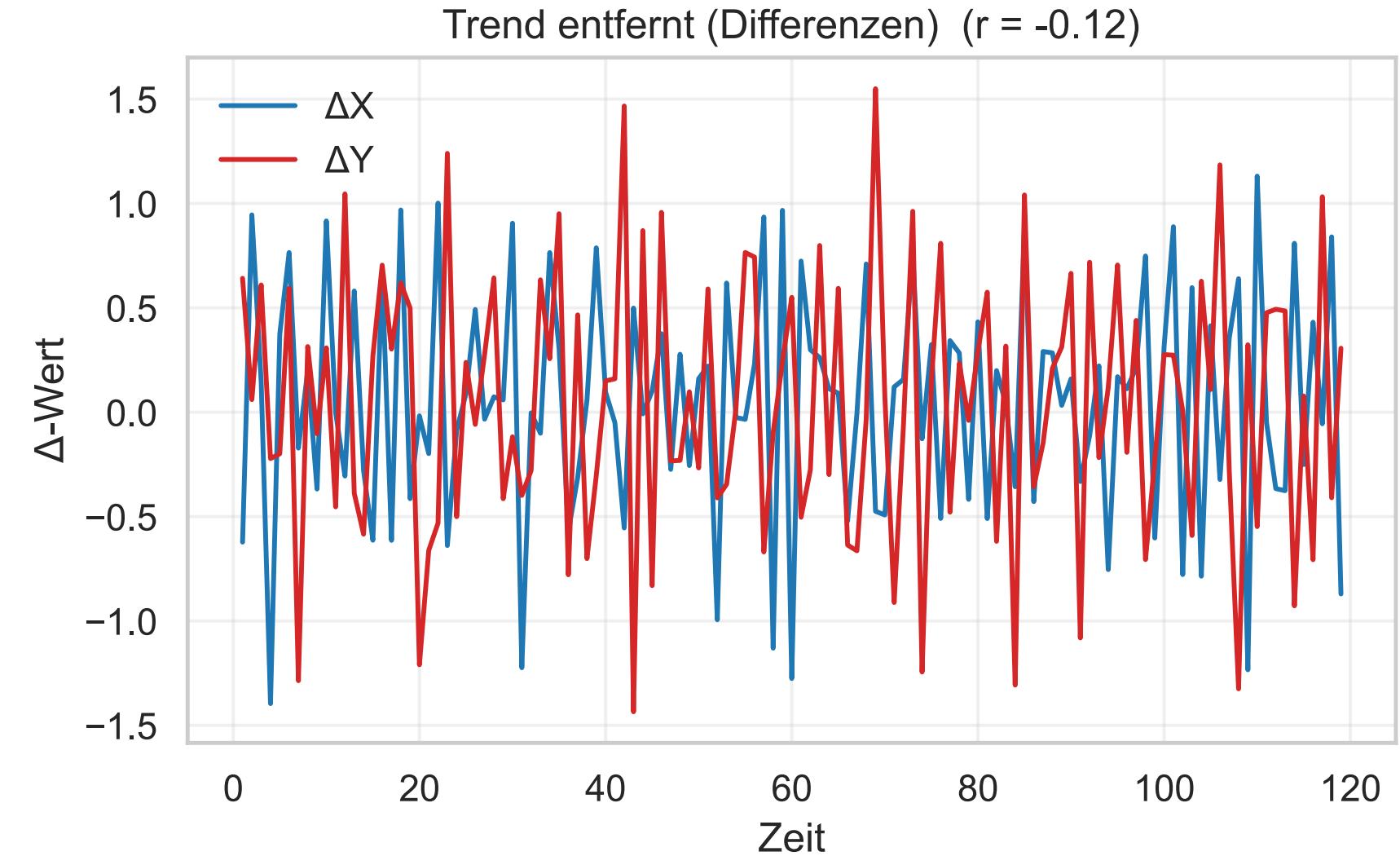
Mini-Check: Wie kann man eine Scheinkorrelation entlarven?

Scheinkorrelation: Hohe r-Werte durch gemeinsame Trends

Parallele Trends ($r = 0.90$)



Trend entfernt (Differenzen) ($r = -0.12$)



Konfundierung: Der Begriff

Eine Drittvariable (Z) verzerrt den scheinbaren Zusammenhang zwischen X und Y .

- Definition: Z beeinflusst **sowohl X als auch Y** .
- Ohne Kontrolle von $Z \Rightarrow$ **Scheinkorrelation**.
- Beispiel: Alter (Z) verzerrt den Zusammenhang: Sport (X) v Gesundheit (Y).
- Die Lösung: **Stratifizieren** (Gruppen bilden) oder **partielle Korrelation**.

Mini-Check: Wie würde man Z statistisch kontrollieren?

Beispiel: Versteckte Variable

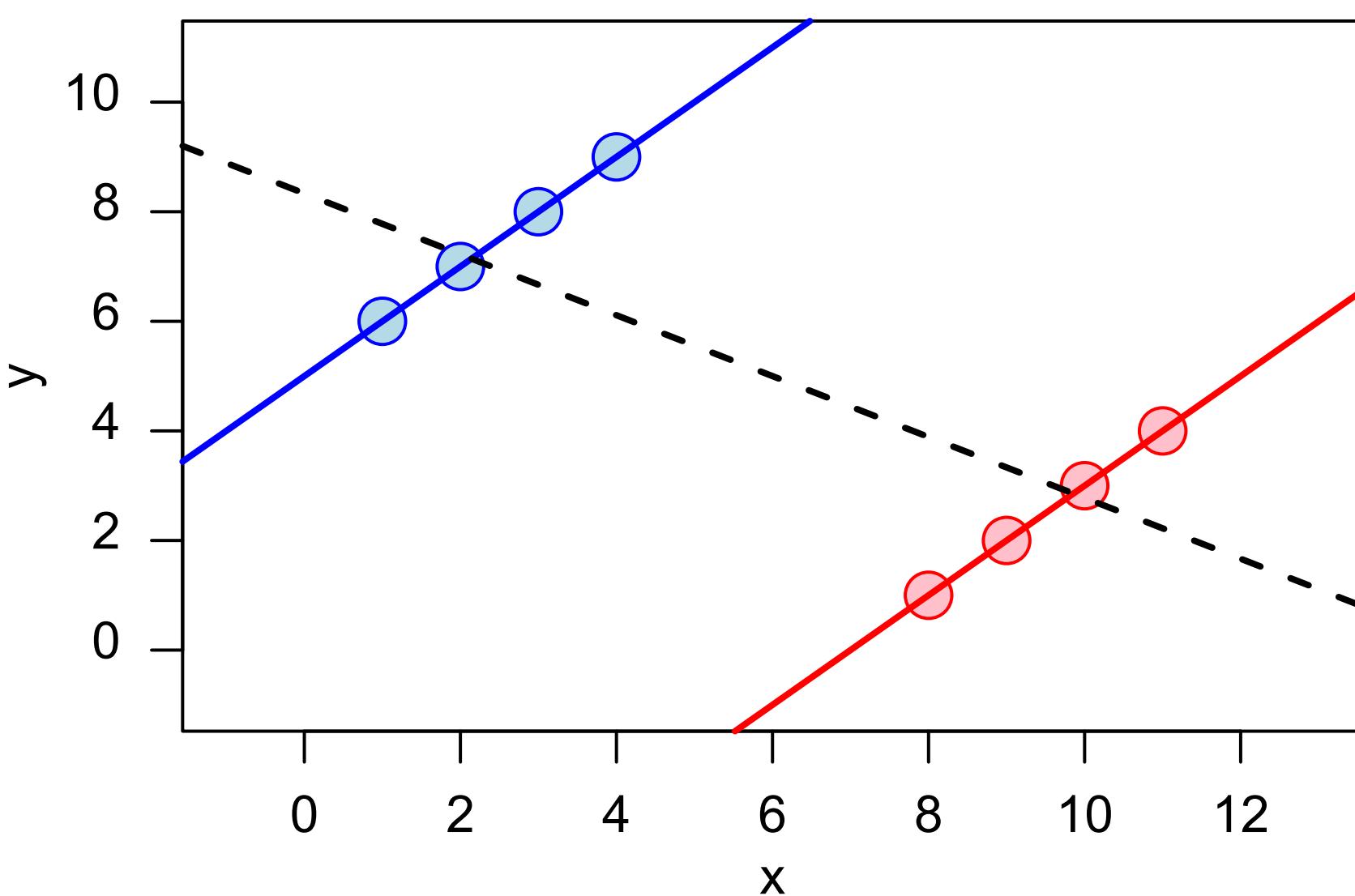
Eine Drittvariable kann den Effekt nicht nur erklären, sondern sogar umkehren.

- X = Kaffeekonsum, Y = Herzerkrankung, Z = Rauchen.
- Ohne Z (aggregiert): Kaffee $\uparrow \Rightarrow$ Herzerkrankung \uparrow .
- Mit Z (kontrolliert): Kaffee \uparrow nur bei Rauchern, aber Kaffee selbst ist neutral.
- Fazit: Z (Rauchen) verursacht die Scheinkorrelation.

Mini-Check: Welche Variable musste man hier «kontrollieren»?

Simpson-Paradox: Einführung

Ein scheinbarer Widerspruch in der Statistik: Ein Trend, der in mehreren Gruppen auftritt, verschwindet oder kehrt sich um, wenn die Gruppen zusammengeführt werden.



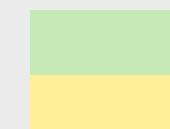
Simpson-Paradox: UC Berkeley (1)

	All		Men		Women	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	12,763	41%	8,442	44%	4,321	35%

Die Zulassungszahlen in Berkeley, zeigten, dass männliche Bewerber mit höherer Wahrscheinlichkeit zugelassen wurden als weibliche.

Simpson-Paradox: UC Berkeley (2)

Dept.	All		Men		Women	
	App.	Adm.	App.	Adm.	App.	Adm.
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%
Total	4526	39%	2691	45%	1835	30%



höhere Erfolgsquote
mehr Bewerbungen

Die bereinigten Daten ergaben dahingegen einen kleinen, aber statistisch signifikanten Vorteil zugunsten von Frauen

Simpson-Paradox: Einführung

Aggregierte Daten können andere Trends zeigen als Subgruppen.

Das ist der Worst Case.

- Definition: Der Trend kehrt sich um, wenn man Gruppen mischt.
- Die Gefahr: falsche Schlüsse bei aggregierten Analysen.

Mini-Check: Welche Information fehlt bei Aggregaten?

Python-Demo: Simpson im Datensatz?

```
tips = sns.load_dataset("tips")
print("Gesamt Korrelation:")
print(tips[["total_bill","tip"]].corr())

print("\nKorrelation pro Geschlecht:")
print(tips.groupby("sex")[["total_bill","tip"]].corr())
```

Python-Demo: Simpson im Datensatz?

Gesamt Korrelation:

	total_bill	tip
total_bill	1.000000	0.675734
tip	0.675734	1.000000

Korrelation pro Geschlecht:

		total_bill	tip
sex			
Male	total_bill	1.000000	0.669753
	tip	0.669753	1.000000
Female	total_bill	1.000000	0.682999
	tip	0.682999	1.000000

Umgang mit Simpson-Effekt

Stratifizierte Analysen vermeiden Fehlschlüsse.

- Strategie 1: Daten nach Konfundierer aufsplitten und separate Analysen pro Gruppe durchführen (Stratifizierung).
- Strategie 2: Partielle Korrelation (lineares Herausrechnen des Z-Einflusses).
- Strategie 3: Regressionsmodelle mit Kontrollvariablen.
- Prinzip: Kontrolliere auf Drittvariablen, bevor du interpretierst.

Mini-Check: Wann würdest du statt Korrelation eine Regression nutzen?

Reflexion: Gefährliche Korrelationen

Interpretation braucht **Skepsis** und **Kontextwissen**.

- Prüfe immer: Mechanismus? Zeitverlauf? Drittvariablen (Z)?
- Hinterfrage Datenquelle und Skalierung. Wer hat die Daten gesammelt?
- **Diskutiere im Team:** Was könnte falsch laufen? (Das ist deine beste Fehlerprüfung!)
- Übung: Formuliere eine Presse-Schlagzeile und entkräfte sie (z.B. «Kaffee macht reich»).

Mini-Check: Was macht eine Korrelation «gefährlich»?

Motivation: Kontrolle von Drittvariablen

Um echte Zusammenhänge zu erkennen, musst du **Drittvariablen (Z) herausrechnen.**

- Beispiel: Training (X) \leftrightarrow Gesundheit (Y), Alter (Z) als Konfounder.
- Ohne Kontrolle von Z : Du wirst den Effekt des Trainings **überschätzen**.
- Die elegante Lösung: **Partielle Korrelation** ($r_{xy.z}$).
- Das Prinzip ist: Du isolierst den Effekt, den du wirklich sehen willst.

Mini-Check: Wann ist die Kontrolle einer Drittvariable besonders wichtig?

Definition: Partielle Korrelation

Misst den Zusammenhang zwischen X und Y , nachdem der **lineare** Einfluss von Z entfernt wurde.

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$$

- Basiert auf den Pearson-Korrelationen (r_{xy}, r_{xz}, r_{yz}) der drei Variablen.
- $r_{xy.z} \neq r_{xy} \Rightarrow$ zeigt, wie stark Z den ursprünglichen Zusammenhang verzerrt hat.

Mini-Check: Was passiert, wenn r_{xz} oder $r_{yz} = 0$ ist?

Rechenbeispiel: Partielle Korrelation

Angenommen: $r_{XY} = 0.7$, $r_{XZ} = 0.8$, $r_{YZ} = 0.9$.

Eingesetzt:

$$r_{xy.z} = \frac{0.7 - (0.8)(0.9)}{\sqrt{(1 - 0.8^2)(1 - 0.9^2)}} = -0.10$$

Interpretation: Der scheinbar starke positive Zusammenhang ($r = 0.7$) **verschwindet** fast und kehrt sich sogar leicht ins Negative (partiell $r = -0.10$).

Mini-Check: Was war hier die verfälschende Variable?

Python: Partielle Korrelation (1)

```
from pingouin import partial_corr

df = sns.load_dataset("penguins")[["bill_length_mm", "bill_depth_mm", "species"]].dropna()

sp_dummies = pd.get_dummies(df["species"], prefix="sp", drop_first=True)

df_num = pd.concat([df[["bill_length_mm", "bill_depth_mm"]], sp_dummies], axis=1)

r_partiell = partial_corr(
    data=df_num,
    x="bill_length_mm",
    y="bill_depth_mm",
    covar=sp_dummies.columns.tolist(),
    method="pearson"
)

print(r_partiell)
```

Python: Partielle Korrelation (2)

```
# 2) Kategoriale Kovariate 'species' in numerische Dummies umwandeln  
#     drop_first=True vermeidet perfekte Multikollinearität  
sp_dummies = pd.get_dummies(df["species"], prefix="sp", drop_first=True)
```

- Kontrolle auf «species» entfernt Artefakte, die nur durch die Grösse der Pinguin-Arten entstehen.
- Vergleich das Ergebnis mit df.corr() (gesamter Datensatz).

Mini-Check: Warum ist Kontrolle auf «species» sinnvoll?

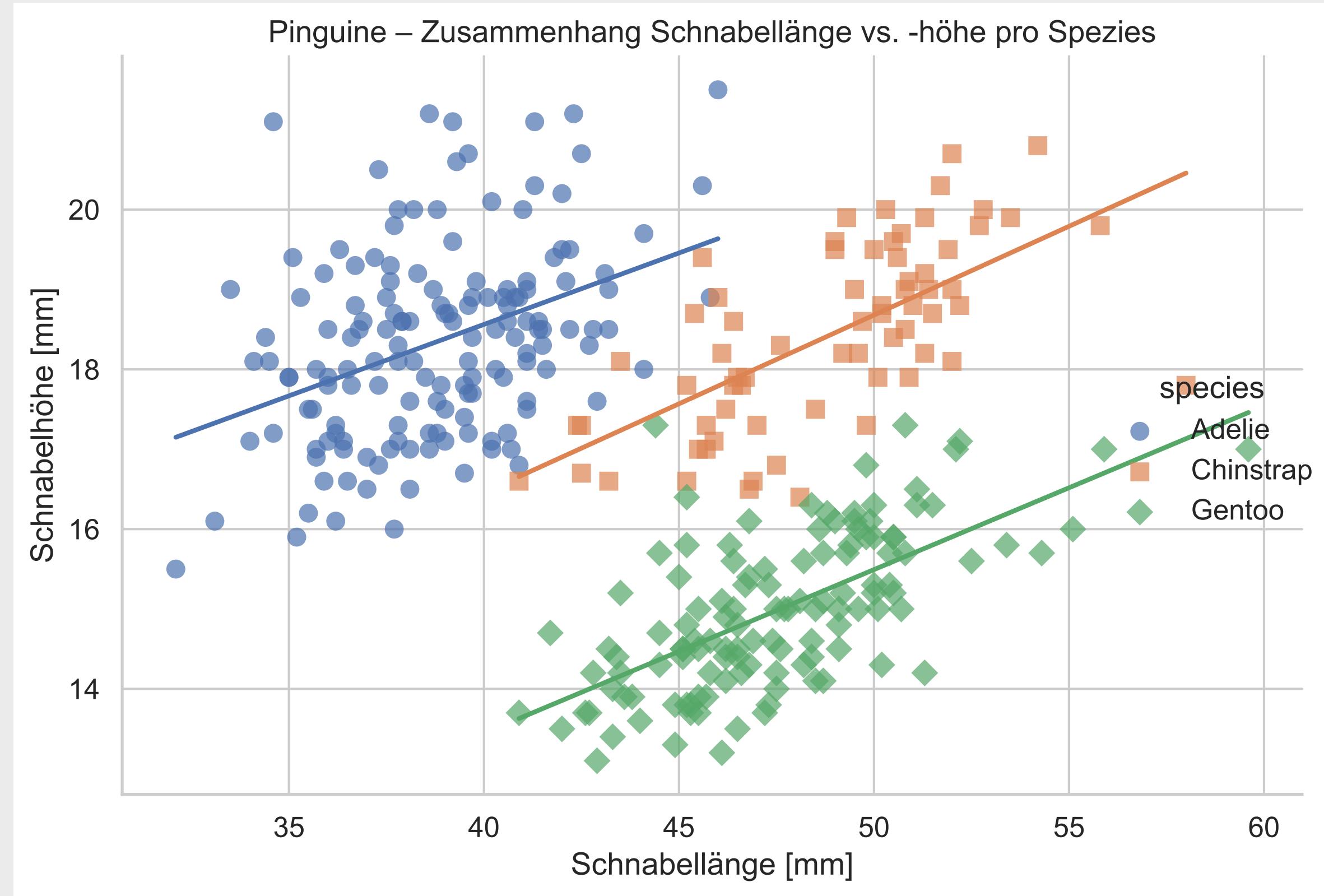
Python: Partielle Korrelation (3)

☰ Partielle Korrelation (kontrolliert für 'species') ☰

	n	r	CI95%	p-val
pearson	342	0.527878	[0.45, 0.6]	8.661124e-26

☰ Gesamtkorrelation (df.corr()) ☰

	bill_length_mm	bill_depth_mm
bill_length_mm	1.000000	-0.235053
bill_depth_mm	-0.235053	1.000000



Interpretation: Was zeigt partielle Korrelation wirklich?

Sie zeigt, was nach Ausschluss gemeinsamer Ursachen übrig bleibt: den reinen Kern.

- $r_{xy.z}$ misst den **reinen, direkten** Zusammenhang.
- Ist $r_{xy.z}$ klein oder null, war Z dominant (Scheinkorrelation).
- Es ist kein Ersatz für Kausalität, aber das beste **Diagnosetool** in der Korrelationsanalyse.
- Es ist die Brücke zur **multiplen Regression**.

Mini-Check: Wann wäre $r_{xy.z} = 0$, obwohl $r_{xy} \neq 0$?

Beispiel: Informatikprojekt

Auch Softwaremetriken zeigen Scheinkorrelationen.

- X = Codezeilen, Y = Bugs, Z = Erfahrung des Entwicklers.
- Ohne Kontrolle: r_{XY} ist positiv (mehr Code \Rightarrow mehr Bugs).
- Mit Kontrolle auf Erfahrung (Z): $r_{XY.Z} \approx 0$.
- Interpretation: **Erfahrung (Z) erklärt beides**. Weniger Erfahrene schreiben mehr Code (langsamer) und machen mehr Fehler.

Mini-Check: Was wäre eine bessere Metrik für Qualität?

Beispiel: Gesundheit & Ernährung

- X = Gemüseverzehr, Y = Cholesterin, Z = Einkommen.
- Ohne Kontrolle: $X \uparrow \Rightarrow Y \downarrow$ (gute Korrelation).
- Mit Kontrolle auf Z (Einkommen): Der Effekt schrumpft.
- Interpretation: Einkommen beeinflusst Ernährung und Gesundheitsversorgung.

Mini-Check: Welche Variable ist hier Konfounder (Z)?

Reflexion: Anwendung auf eure Projekte

Korrelation ist nur der erste Schritt deiner wissenschaftlichen Analyse.

- **Denk nach:** Welche Variablen in deinem Projekt sind potenziell korreliert?
- Prüfe: Linear oder monoton? Gibt es Drittvariablen (Z)?
- Nutze Heatmap + Scatter zur Exploration.
- **Diskutiere im Team:** Wo könnten Scheinkorrelationen auftreten? (Das ist die beste Fehlerprüfung.)

Mini-Check: Welche Korrelation würdest du am kritischsten prüfen?

Key Takeaways: Praxisfallen

- Korrelation \neq Kausalität: r ist nur ein Hinweis, kein Beweis! → immer experimentelle Kontrolle.
 - Konfundierung (Z): Die häufigste Falle! Drittvariable erzeugt Scheinkorrelation.
→ partieller Korrelation oder Stratifizierung.
 - Simpson-Paradox: Das Aggregat lügt! → immer Gruppenanteile prüfen.
 - Pflicht: Sei skeptisch!
Plotte, prüfe Konfundierer, prüfe Subgruppen → kausaler Schluss
- 👉 Dein Job ist nicht, Korrelationen zu finden, sondern Scheinkorrelationen zu eliminieren.

Zusammenfassung & Ausblick

Überblick über alle behandelten Masse und Fallstricke.

Der Weg: Kovarianz \Rightarrow Pearson \Rightarrow Spearman/Kendall \Rightarrow partielle Korrelation.

- **Erste Regel:** mit Visualisierung immer prüfen (Plotten ist Pflicht!).
- **Zweite Regel:** Korrelation \neq Kausalität.
- **Dritte Regel:** Drittvariablen (Z) erkennen und kontrollieren (Simpson!).

Mini-Check: Welches Korrelationsmass würdest du wann verwenden? (Lineare Daten vs. Rangdaten)

Quiz - Aktive Wiederholung

Kahoot Quiz VL4: Korrelation und Zusammenhangsanalyse

Wahrscheinlichkeiten & Verteilungen

Nach der Analyse von Zusammenhängen folgt die Modellierung von **Zufall**.

- **Korrelation** ⇒ liefert dir die erste Idee von Abhängigkeit.
- **Wahrscheinlichkeit** ⇒ die formale Grundlage für die Unsicherheit und Signifikanz.
- Nächste Sitzung: **Bernoulli, Binomial, Poisson, LLN**.
- **Deine Vorbereitung:** PSDS Kapitel 3 lesen.

Mini-Check: Was ist der Unterschied zwischen Korrelation und Wahrscheinlichkeit?