Foto: Pauline Loroy auf Unsplash

Principal Components Analysis

… und ihr Einsatz bei der Clusteranalyse

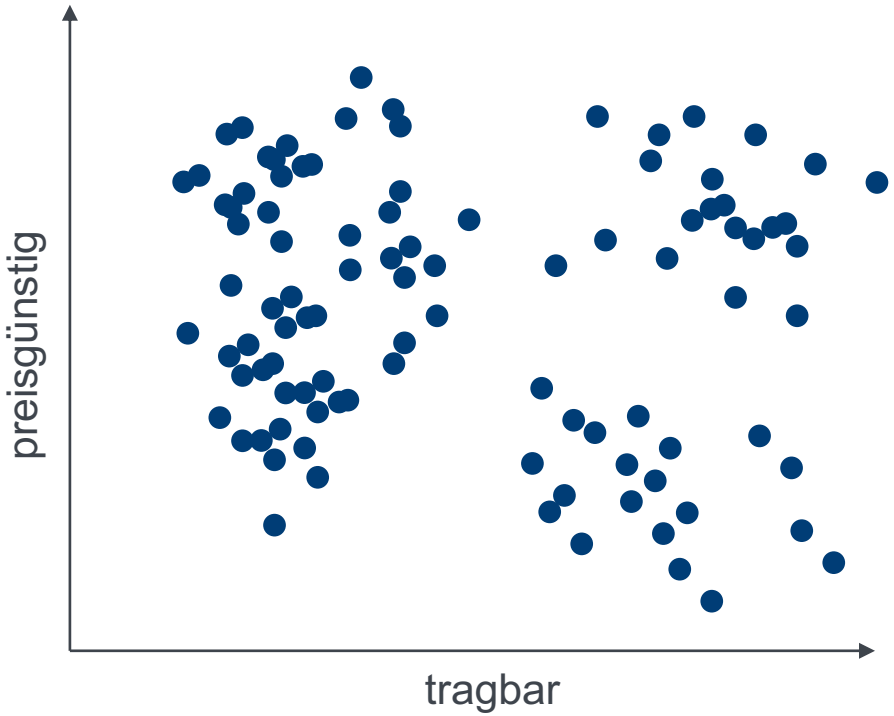
INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

INNOVET
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Clustering mehrdimensionaler Daten

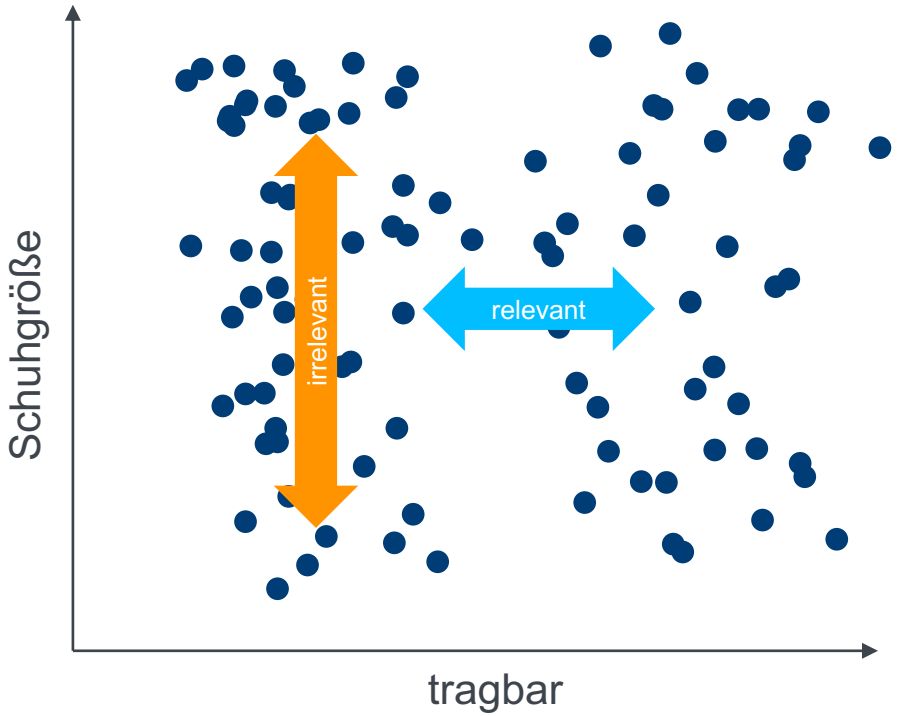
Clustering sinnvoller Dimensionen



2 Gruppen von Kund*innen: Tragbarkeit der Sitze entweder eher wichtig oder eher unwichtig

Tragbarkeit als Dimension sinnvoll

Clustering mit weniger sinnvollen Dimensionen

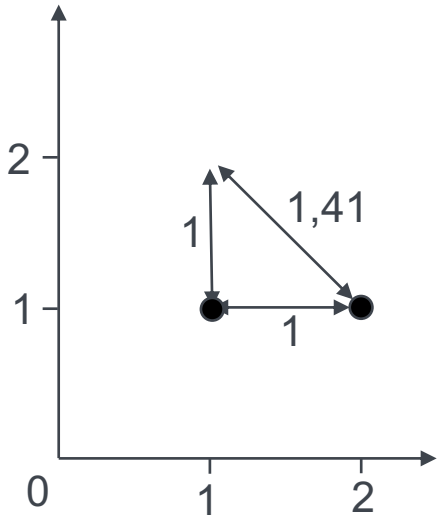


2 Gruppen von Kund*innen: Tragbarkeit der Sitze entweder eher wichtig oder eher unwichtig

Würde vermutlich beim Clustern nicht mehr erkannt

Irrelevante Dimensionen können bei der Clusteranalyse interessante Strukturen „verschleiern“.

Anders erklärt



- Beispiel:
  - 2 Punkte, Differenz in allen Dimensionen: 1
  - Sinnvolle Dimensionen: nur eine
- Euklidische Distanz bei…
  - 1 Dimension: 1
  - 2 Dimensionen: 1,41
  - 3 Dimensionen: 1,73
- Differenz von 1 in der sinnvollen Dimension im Verhältnis zur Gesamtdistanz mit jeder neuen Dimension kleiner!

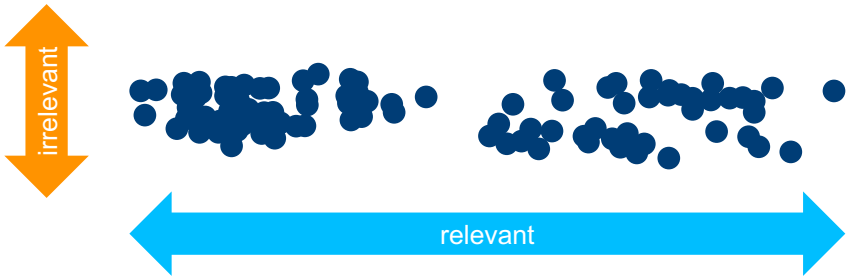
Take-home Message

Bei der Clusteranalyse sollte man sich gut überlegen, welche Dimensionen relevant sein müssten!

- Hier: welche Angaben der Kund*innen könnten helfen, sinnvolle Gruppen von Kundentypen zu finden?
- Sinnvolle Gruppen – hier z.B. Gruppen, die vermutlich unterschiedliche Produkte kaufen werden

Dimensionalitätsreduktion

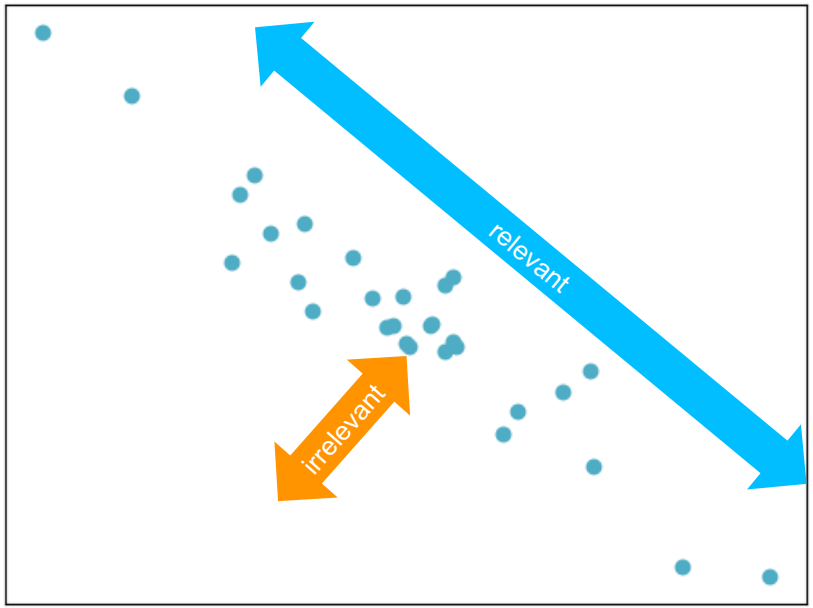
Ein Beispiel



- Dimension x: hohe Varianz
  - daher vermutlich relevant
- Dimension y: weniger Varianz
  - daher vermutlich weniger relevant

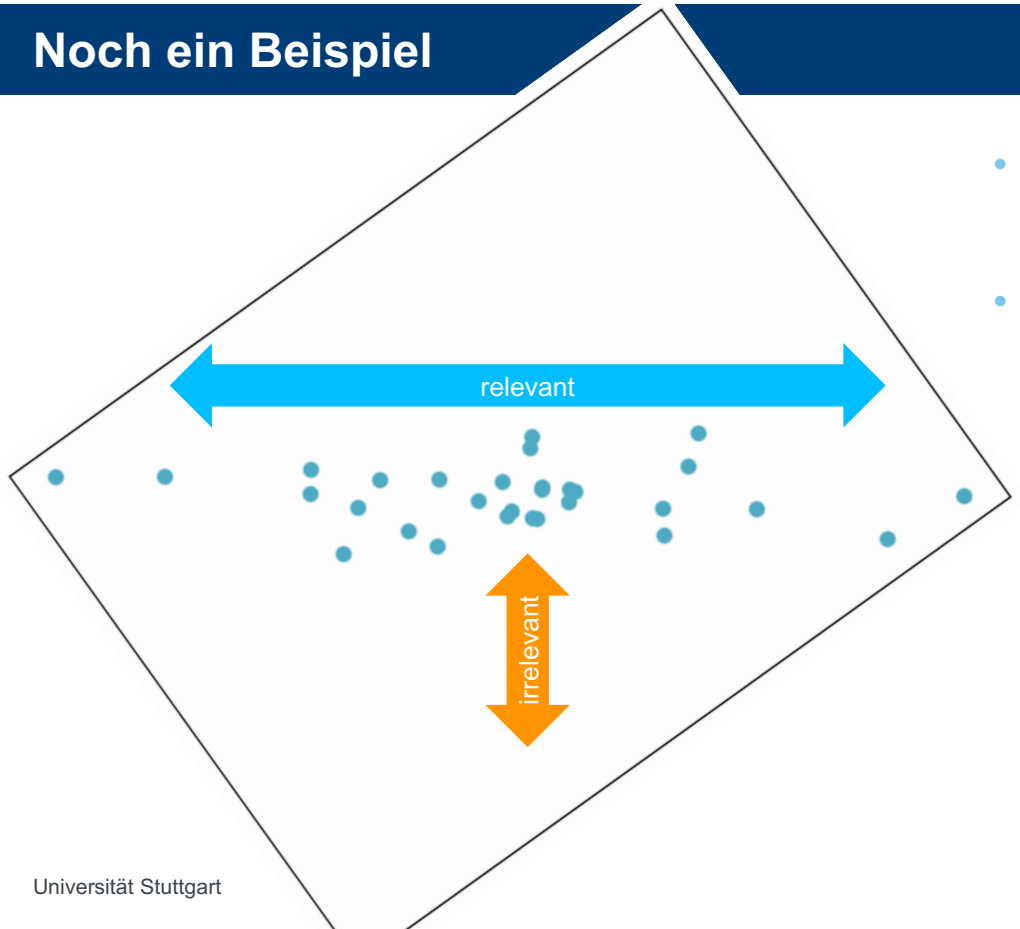
Idee: die weniger relevante Dimension weglassen!

Noch ein Beispiel



- Welche Dimension ist hier relevanter?? x oder y?
- Idee: das Koordinatensystem drehen!

Noch ein Beispiel



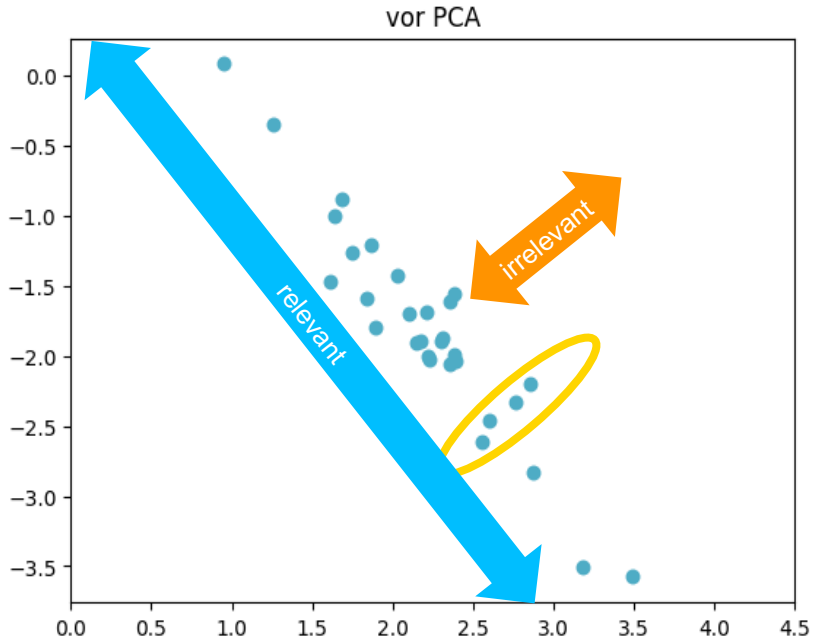
- Welche Dimension ist hier relevanter?? x oder y?
- Idee: das Koordinatensystem drehen!

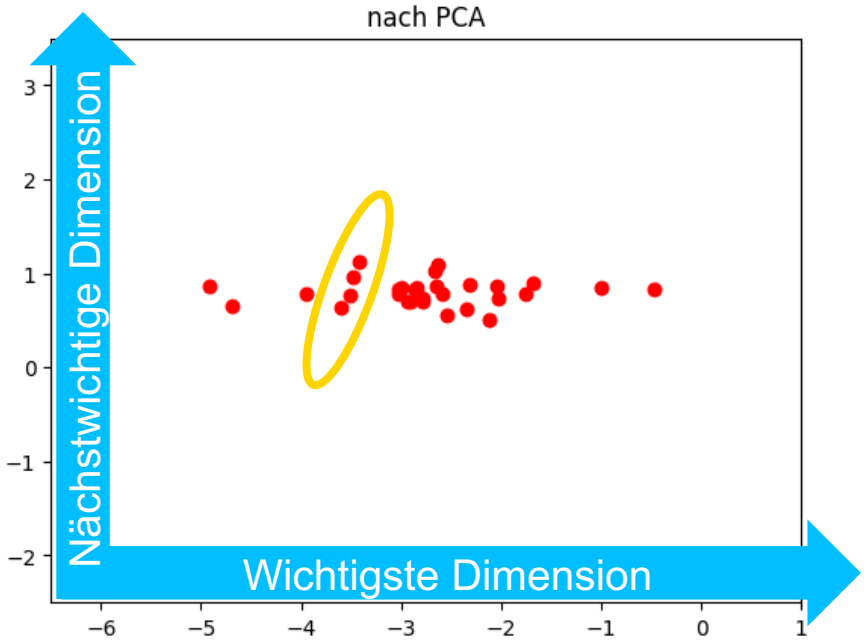
PCA-Analyse
„Principal Components Analysis“
= Hauptkomponentenanalyse

PCA steht für Principal Components Analysis (Hauptkomponentenanalyse).

Die Grundannahme der PCA ist, dass Dimensionen, in denen die Daten am meisten Varianz aufweisen, am relevantesten sind.

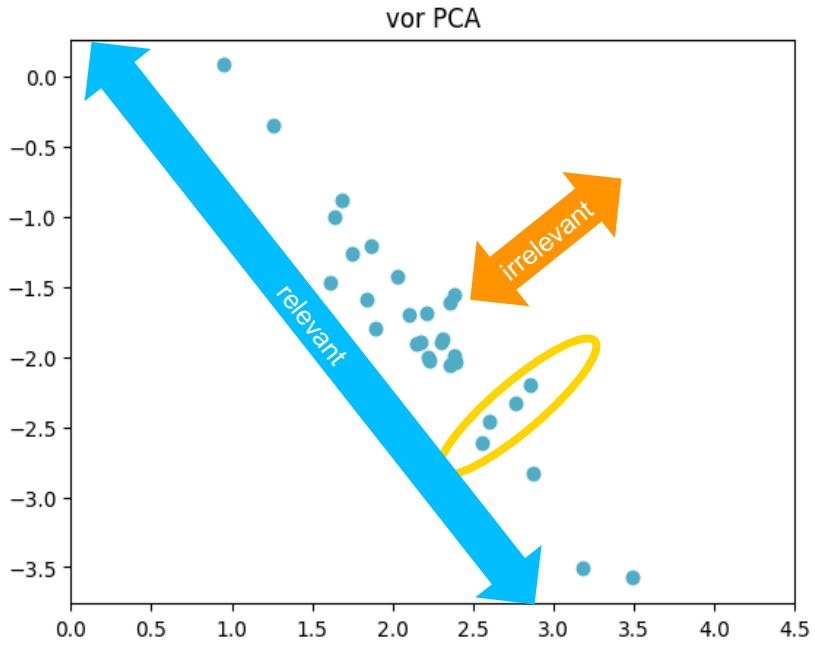
PCA-Analyse für unser Beispiel

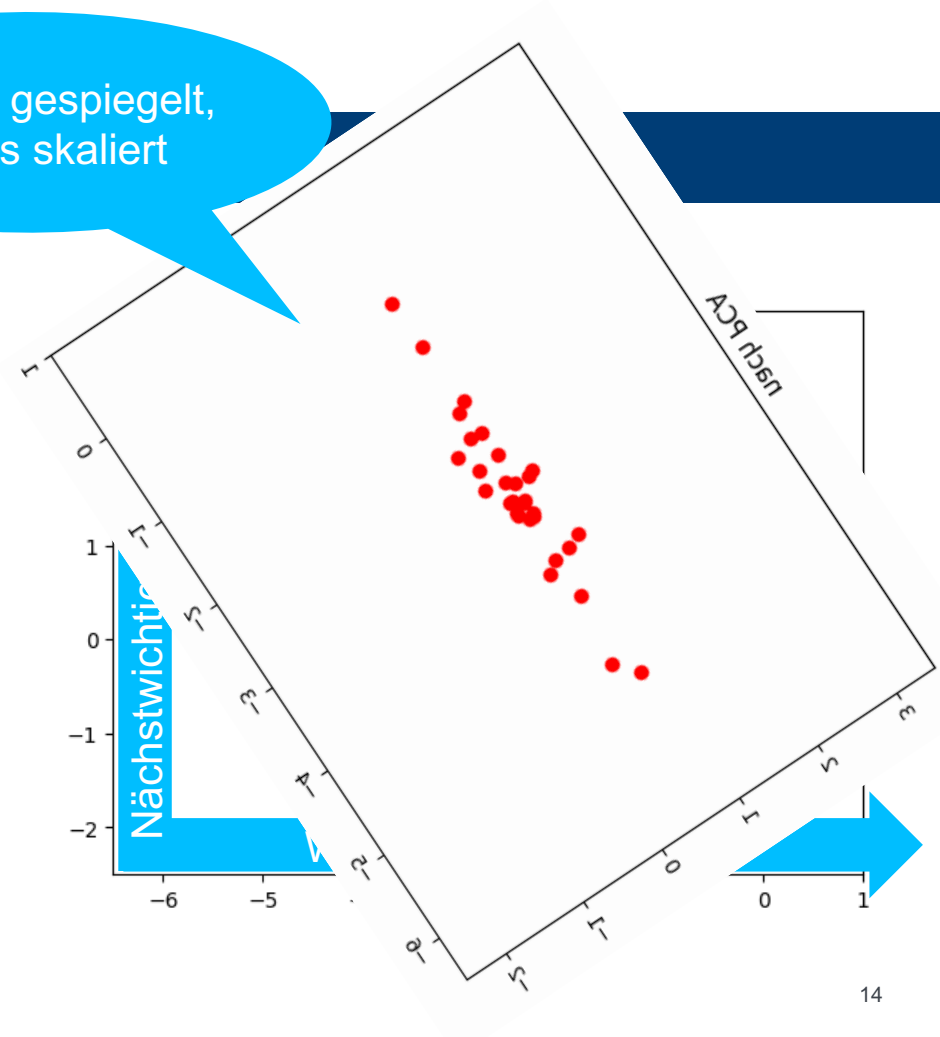




Transformation in unserem B

gedreht, gespiegelt,
anders skaliert

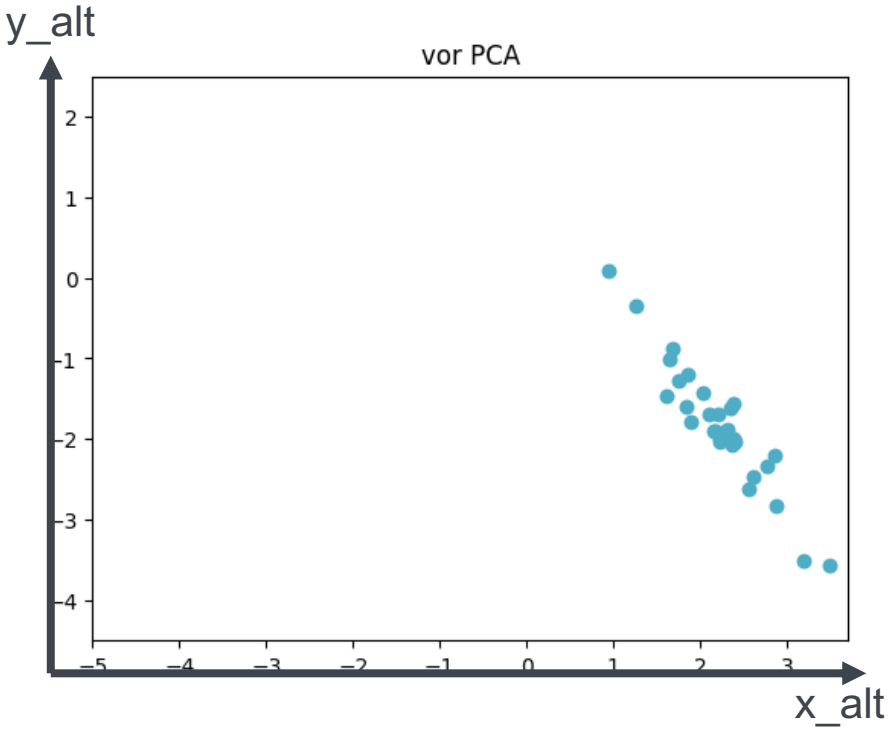


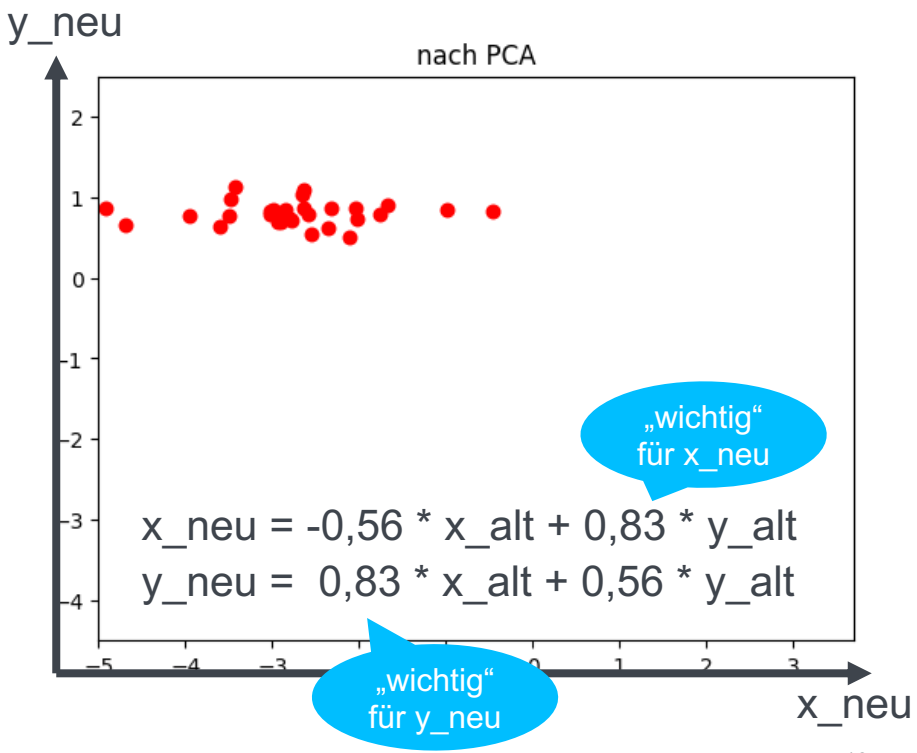


Die PCA transformiert die Daten so, dass die neuen Dimensionen nach Relevanz sortiert sind.

Dadurch kann man höhere Dimensionen eher weglassen.

Vorher-nachher im selben Koordinatensystem





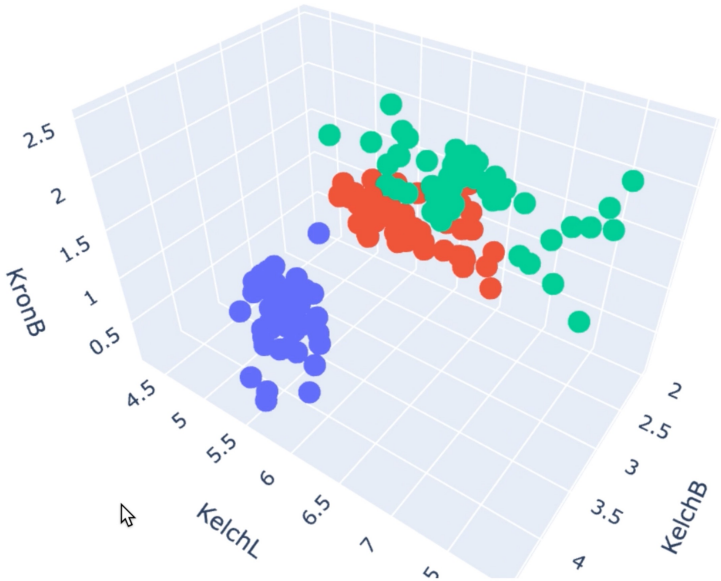
Die Koeffizienten der PCA-Transformation geben Aufschluss darüber, welche der ursprünglichen Dimensionen wie viel zu den neuen Dimensionen beitragen.

Koeffizienten mit großem Betrag deuten an, dass die Dimension viel beiträgt.

Visualisierung von Ergebnissen bei der Clusteranalyse

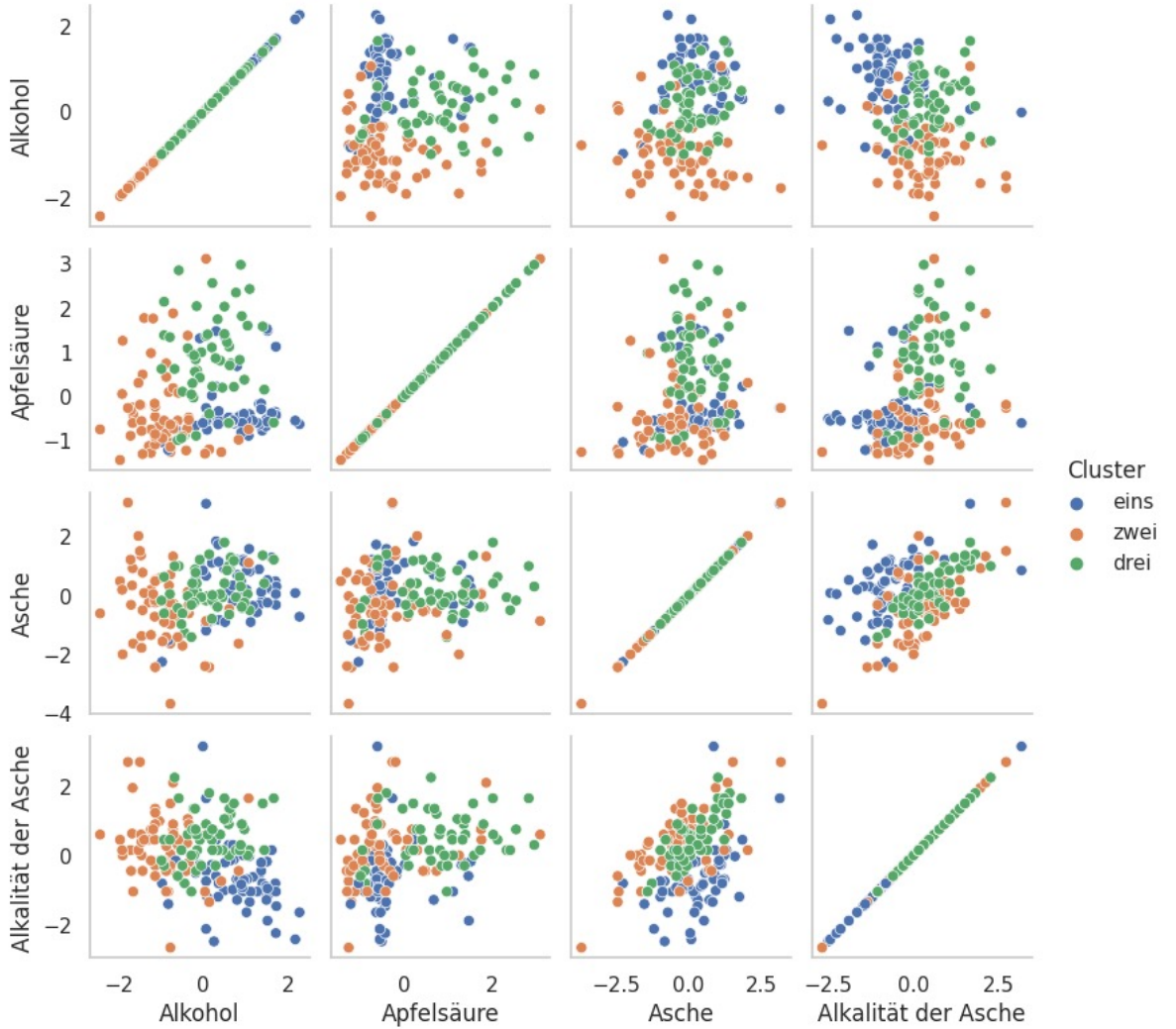
Visualisierung von mehrdimensionalen Daten

- Zweidimensionale Daten
  - Normale Grafik, kein Problem!
- Dreidimensionale Daten
  - Als 3D-Grafik
  - Besser als Film



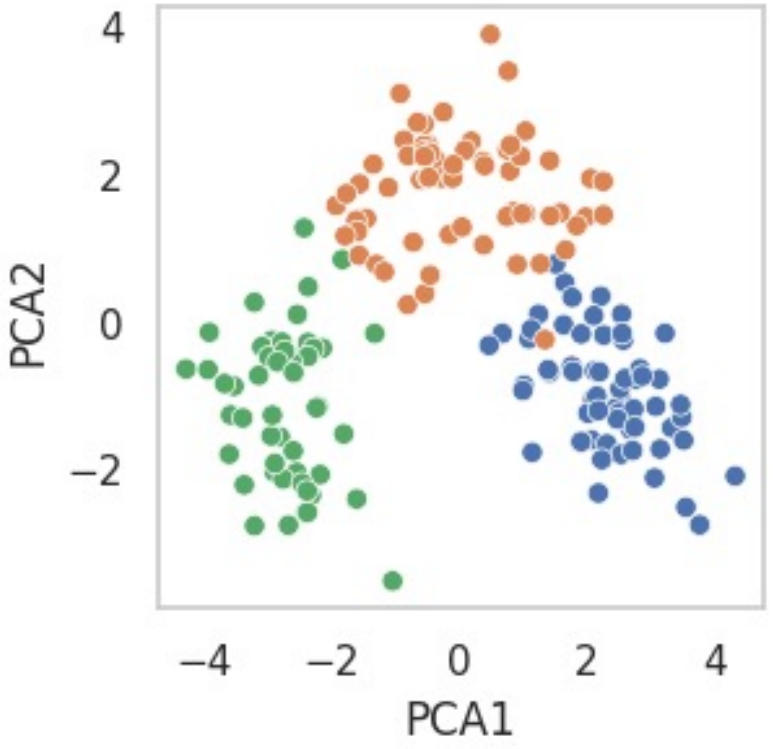
Visualisierung von mehrdimen

- Und bei mehr Dimensionen???
- Beispiel Clusterergebnisse der Weindaten aus Jupyter Notebook
- Hier: 4 Dimensionen, paarweise im Scatterplot



PCA zur Visualisierung

- Statt dessen: zur Visualisierung mit PCA auf 2 Dimensionen reduziert
- (Clusteranalyse auf allen Dimensionen)



Die PCA eignet sich auch dazu, mehrdimensionale Daten zur zweidimensionalen Visualisierung auf zwei Dimensionen zu reduzieren.

INNOVET

KI B³

Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM

INNOVET

Bundesministerium für Bildung und Forschung

bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

Quellenverzeichnis

Titelfoto: Pauline Loroy (https://unsplash.com/de/@paulinel), Weiß-rotes schiefes Gebäude unter blauem Himmel, auf Unsplash (https://unsplash.com/de/fotos/weiss-rotes-schiefes-gebaude-unter-blauem-himmel-EIJ3ogh6EVk), lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.