

# Clustering

mit Scikit-Learn

# Agenda

1. Definition Cluster-Analyse → Marius
2. Kontext Datensatz → Marius
3. Daten aufbereiten → Mario
4. Daten visualisieren → Mario
5. Cluster-Analyse: kMeans → Marius
6. Cluster-Analyse: Hierarchisch → Mario
7. Kritische Reflexion → Marius
8. Fazit → Mario

# Definition Cluster-Analyse

- Verfahren des maschinellen Lernens
- In einer Menge von **Daten** „ähnliche“ **Gruppierungen (Cluster)** erkennen
- Einsatz unterschiedlicher **Algorithmen** zur Bildung der Cluster

# Kontext Datensatz

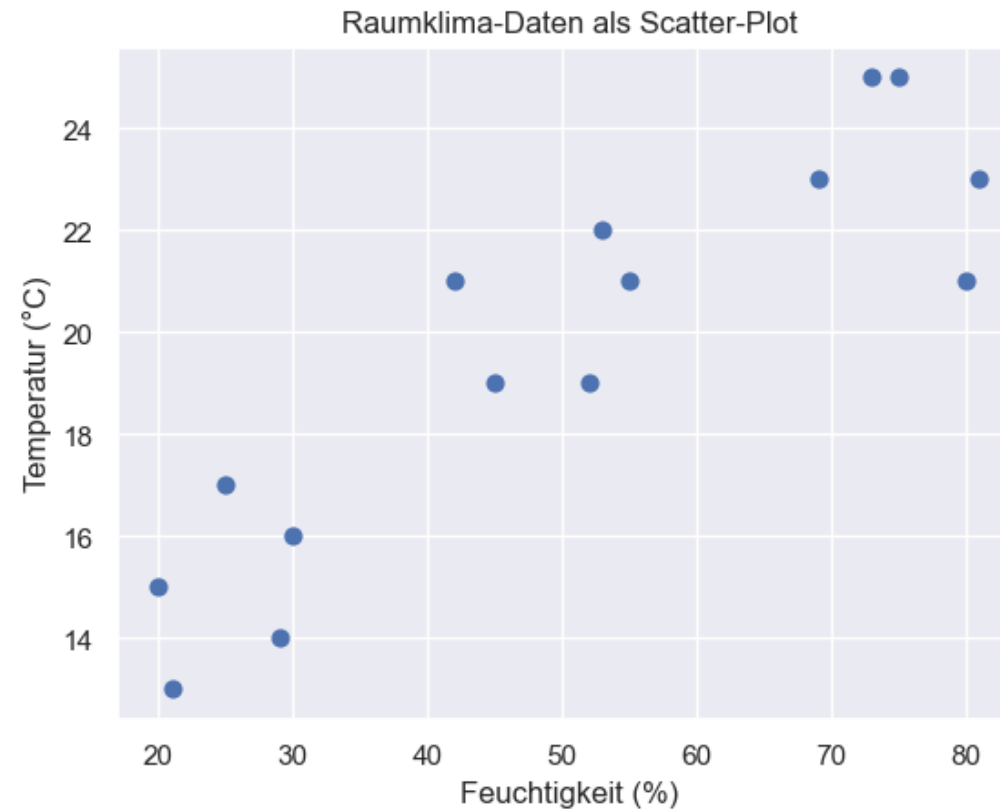
- **Raumklima**-Datensatz
- 15 Messungen mit Temperatur (°C) und Luftfeuchtigkeit (%)
- Unterschiedliche Kombinationen und damit Klima-Arten
- Gibt es ein **optimales Klima**?

# Daten aufbereiten

- In der Regel: **Dubletten** entfernen, **Metriken** anpassen, Umgang **Nullwerte**, usw.
- In diesem Fall: Nicht notwendig

Feuchte (in %)	Temperatur (in °C)
42	21
45	19
52	19
55	21
53	22
69	23
80	21
81	23
73	25
75	25

# Daten visualisieren



Visualisiert mit Matplotlib

# Cluster-Analyse: kMeans

- Lorem

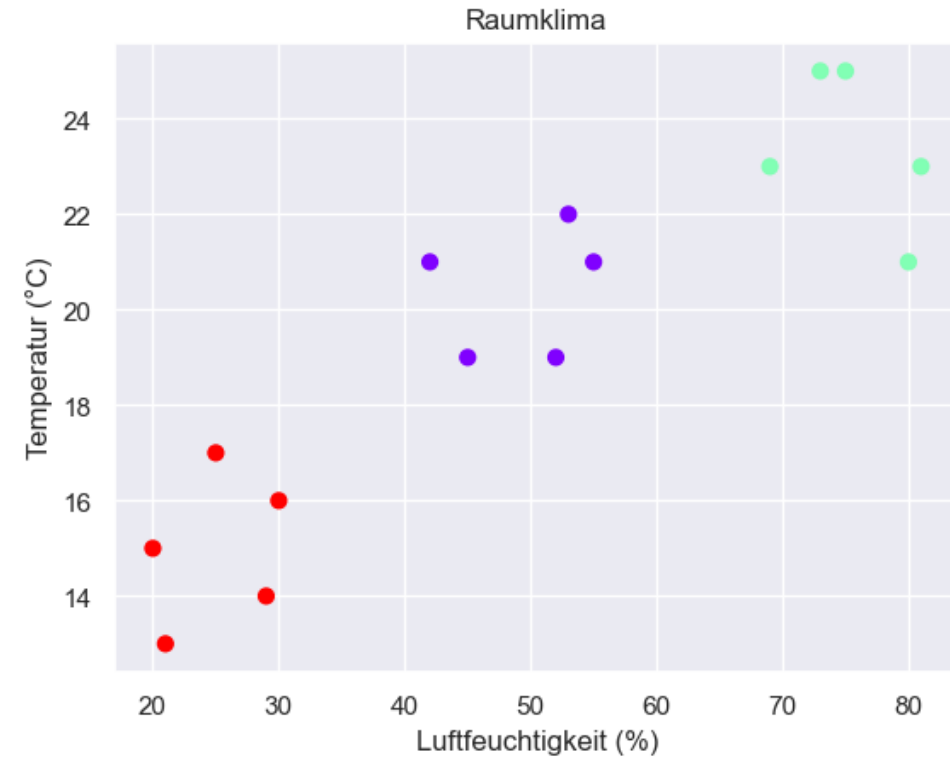
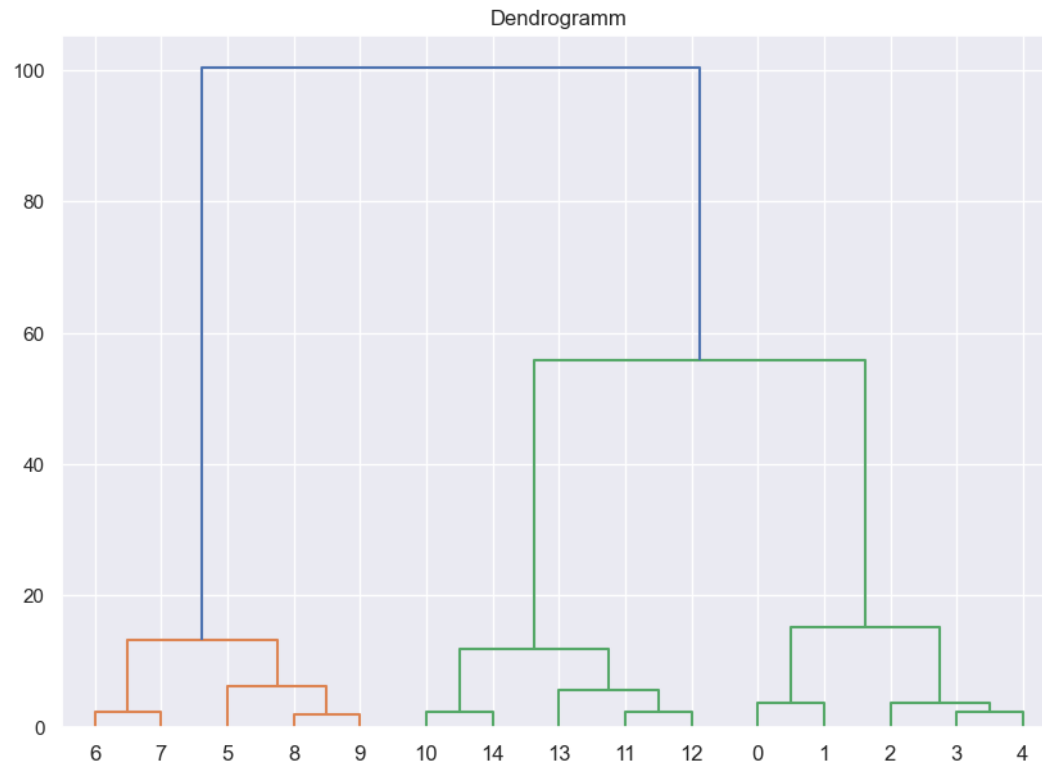
# Cluster-Analyse: Hierarchisch

- **Agglomerative** Cluster-Analyse
- Darstellung in **Dendogramm**
- Abstandfunktion: **Euklidische** Distanz
- Fusionsvorschrift: **Ward** Methode



# Cluster-Analyse: Hierarchisch

Visualisiert mit Matplotlib



# Ausblick

- Anzahl Features (d):  $> 2$
- Hyperparameter-Tuning
- Bias
- Over- & Underfitting
- Vergleich der unterschiedlichen Cluster-Scores