

## Unsupervised Learning – PCA und k-Means Clustering

### Datensatz: Country-data.csv

Dieser Datensatz enthält Daten über die sozio-ökonomischen und Gesundheits-Faktoren von über 160 Ländern. Dazu gehören Angaben zur Kindersterblichkeit, Export- und Import-Werte pro Kopf und Ausgaben für das Gesundheitswesen. Diese Daten wurden für HELP International (internationale humanitäre NGO) erhoben, um Länder hinsichtlich ihres Entwicklungsstandes zu kategorisieren.

**Datenquelle:** [www.kaggle.com/kmeans-on-country-data](https://www.kaggle.com/kmeans-on-country-data)

### Aufgaben:

- 1) Untersuchen Sie den Datensatz Country-data.csv hinsichtlich der oben genannten Fragestellung. Die entsprechenden Metadaten zu den gegebenen charakteristischen Länder-Merkmalen finden Sie in der Datei data-dictionary.csv!
  - 2) Verschaffen Sie sich mithilfe der Datenvorbereitung und der Explorativen Datenanalyse EDA einen Überblick über Qualität und Inhalte der Daten! Gibt es Korrelationen zwischen verschiedenen Variablen?
  - 3) Wenden Sie eine Dimensionsreduktion mittels Principal Component Analysis PCA auf die Daten an und entscheiden Sie, wie viele Hauptkomponenten Sie in die weitere Analyse einbeziehen werden!
  - 4) Wenden Sie eine Klassifikation mittels k-Means-Clustering auf den Datensatz an! Wie viele Cluster sind für die weitere Interpretation sinnvoll? Wie interpretieren Sie die gefundenen Clusterzuordnungen?
  - 5) Fassen Sie ihre Ergebnisse kurz zusammen!
- **Abgabe bis Donnerstag 05.12. 16:00 Uhr zwei Folien pro Gruppe (pdf!!), im Namen des Files bitte Gruppennummer bzw. Namen der Verfasser + Nummer der Übung verwenden (z.B. GruppeXX\_uebung-a4.pdf)**
  - **in Verzeichnis Upload\_DASC hochladen:** <https://nc.ufz.de/s/ai2zHScBEic8S8r>  
(Passwort: !DASC\_ba\_2024)
  - Folie 1: Grafik zu Aufgaben 2  
Folie 2: Grafik auswählen aus Aufgabe 3 + 4, kurze Zusammenfassung der Ergebnisse aus der Analyse
  - **5-Minuten Pitch: Jede Gruppe präsentiert kurz ihre Ergebnisse (5 min maximal) zu Beginn der nächsten Veranstaltung am 06.12.24**