

Miniprojekt: Clustering

Szenario

Ein Einkaufszentrum möchte besser verstehen, wer dort einkauft, um auf dieser Basis Kundengruppen bilden und diese dann individueller ansprechen zu können. Dafür wurden von 200 Kunden Daten erhoben zu deren *Geschlecht*, *Alter*, *Jahreseinkommen* (in T\$) und *Ausgabenbereitschaft* (als Wert zwischen 1 und 100). Nun sollen erste Gruppen gebildet und deren Charakteristika herausgearbeitet werden.

Aufgaben

1. Zur Vorbereitung erarbeiten Sie mit Hilfe des Buches Python Machine Learning von Sebastian Raschka (<https://www.packtpub.com/free-ebook/python-machine-learning>) die Themen
 - Encoding und Scaling aus Kapitel 4 (und als weitere Quelle https://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html)
 - Hauptachsenanalyse (Principal Component Analysis (PCA)) aus Kapitel 5

Ferner sollten Sie recherchieren, was der *CRISP-DM-Prozess* ist und sich überlegen, was dieser mit der vorliegenden Aufgabenstellung zu tun haben könnte.

2. Lesen Sie dann die Daten aus der Datei `data/Kunden_EKZ.csv` in ein *pandas Data Frame* innerhalb eines *Jupyter Notebooks* ein. Überlegen Sie dabei, wie mit der Spalte `Customer_ID` umzugehen ist.
3. Führen Sie ein Profiling der eingelesenen Daten durch.
4. Bereiten Sie dann die Daten so auf, dass mit zwei erklärenden Variablen gearbeitet werden kann, die den größten Teil der Varianz erklären.
5. Führen Sie anschließend ein Clustering mit der optimalen Anzahl an Clustern durch. Visualisieren Sie das Ergebnis auch geeignet.
6. Interpretieren Sie daraufhin das Ergebnis, indem Sie die gefundene Kategorisierung auf die Ausgangsdaten anwenden und versuchen diese aufgrund typischer Werte der erfassten Daten zu charakterisieren.
7. Ordnen Sie schließlich Ihre Tätigkeiten zur Lösung des Miniprojekts den Phasen des CRISP-DM zu. Welche Phasen wurden ggfs. noch nicht ausgeführt und wie könnten diese im vorliegenden Fall möglicherweise aussehen?

Bei der Lösung der Aufgabenstellung sollten Sie darauf achten, Ihr Vorgehen in ausreichendem Maße zu erläutern (indem Sie die Markdown-Möglichkeiten des Jupyter Notebooks nutzen).