

Clustering

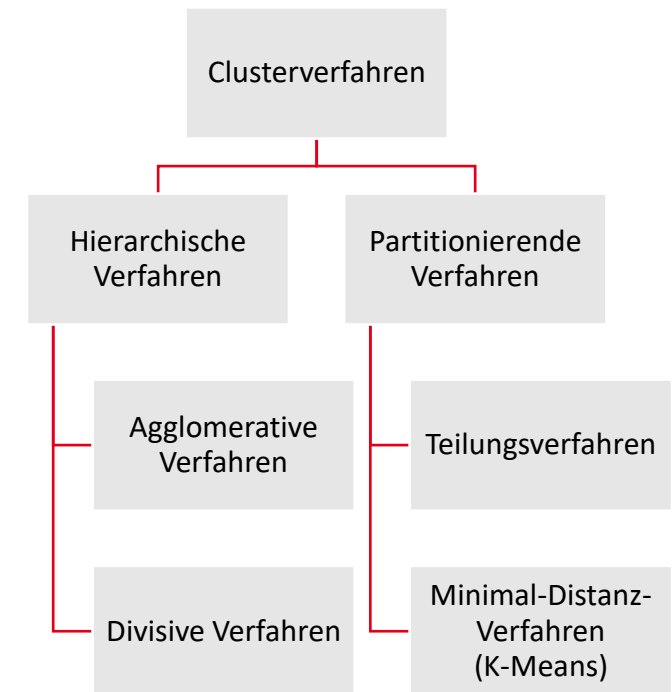
mit Scikit-Learn

Agenda

1. Definition Cluster-Analyse
2. Kontext Datensatz
3. Daten visualisieren & aufbereiten
4. Cluster-Analyse: Hierarchisch
5. Cluster-Analyse: k-Means
6. Fazit
7. Interpretation des Ergebnisses

Definition Cluster-Analyse

- Primäres Ziel clusteranalytischer Auswertungsverfahren ist, eine **Menge von Klassifikationsobjekten** in **homogene Gruppen** (Klassen, Cluster, Typen) **zusammenzufassen**. *(Bacher, J. (2010))*
- Unterscheidung in **hierarchische** und **partitionierende Verfahren**
- **Exploratives Datenanalyseverfahren**
- Anwendung im Bereich des **maschinellen Lernens** (unüberwacht)
- Cluster-Analyse \neq Klassifizierung



Eigene Darstellung nach Backhaus, K (2021)

Kontext Datensatz

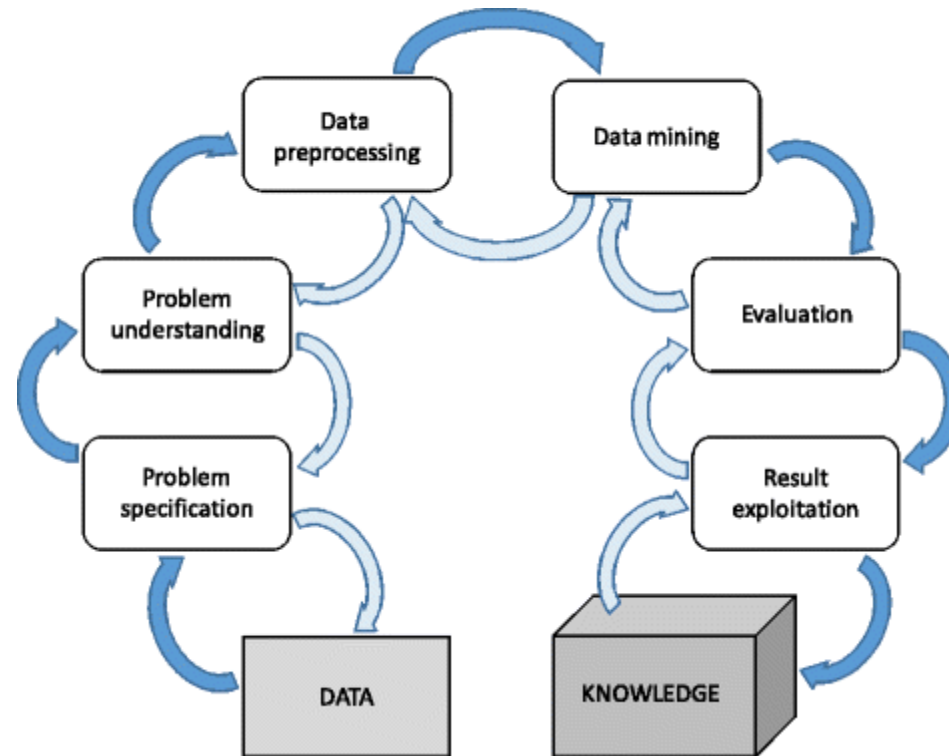
#	Package No	Shipment No	Gross Weight (kg)	Width (cm)	Height (cm)	Length (cm)
0	1007530-2011-03239	1000088	23	35	30	35
1	1007530-2011-03241	1000310	150	60	55	80
2	1007530-2011-03242	1000346	0,5	14	15	19
3	1007530-2011-03243	1000456	1,5	20	20	29
4	1007530-2011-03244	1000796	1	10	10	10
5	1007530-2011-03245	1000957	75	82	81	120
6	1007530-2011-03246	1000957	41	80	34	120
7	1007530-2011-03247	1001184	1.340	220	112	406
8	1007530-2011-03249	1001408	0,5	20	20	29
9	1007530-2011-03250	1001563	5	45	35	45

Original-Datensatz (erste zehn Zeilen)

- **Logistik**-Datensatz
 - Identifier „Package No“
 - Information zu Abmaßen und Gewichten
Kombinationen ergeben verschieden große und schwere Packstücke
 - Zusatzinformation „Shipment No“
- Welche Gruppen gleichartiger Packstücke können gebildet werden, um diese mit spezialisierten Teams zu bearbeiten?

Daten aufbereiten

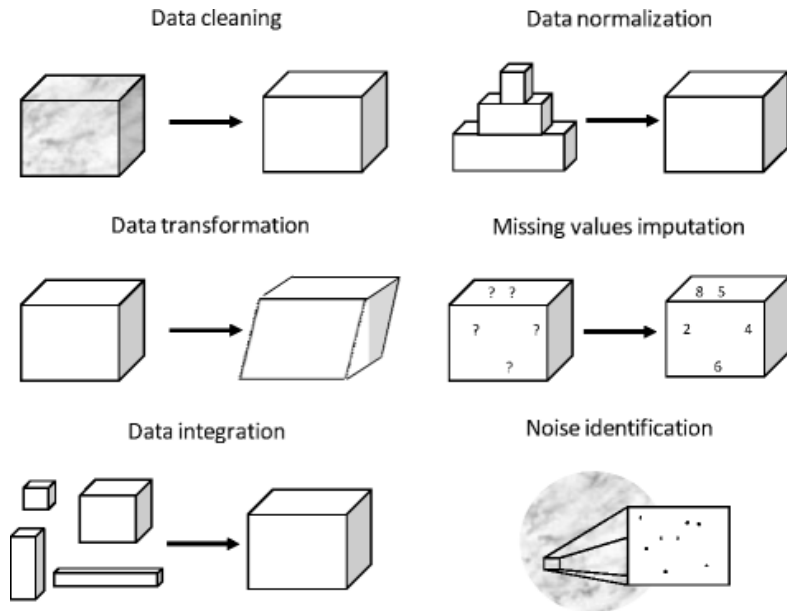
Entnommen aus: García, Salvador u. a. (2016)



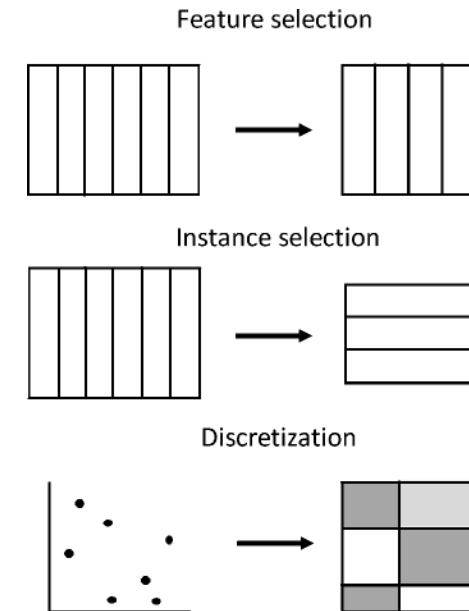
Knowledge Discovery in Databases - Prozess

Daten aufbereiten

Entnommen aus: [García, Salvador u. a. \(2016\)](#)

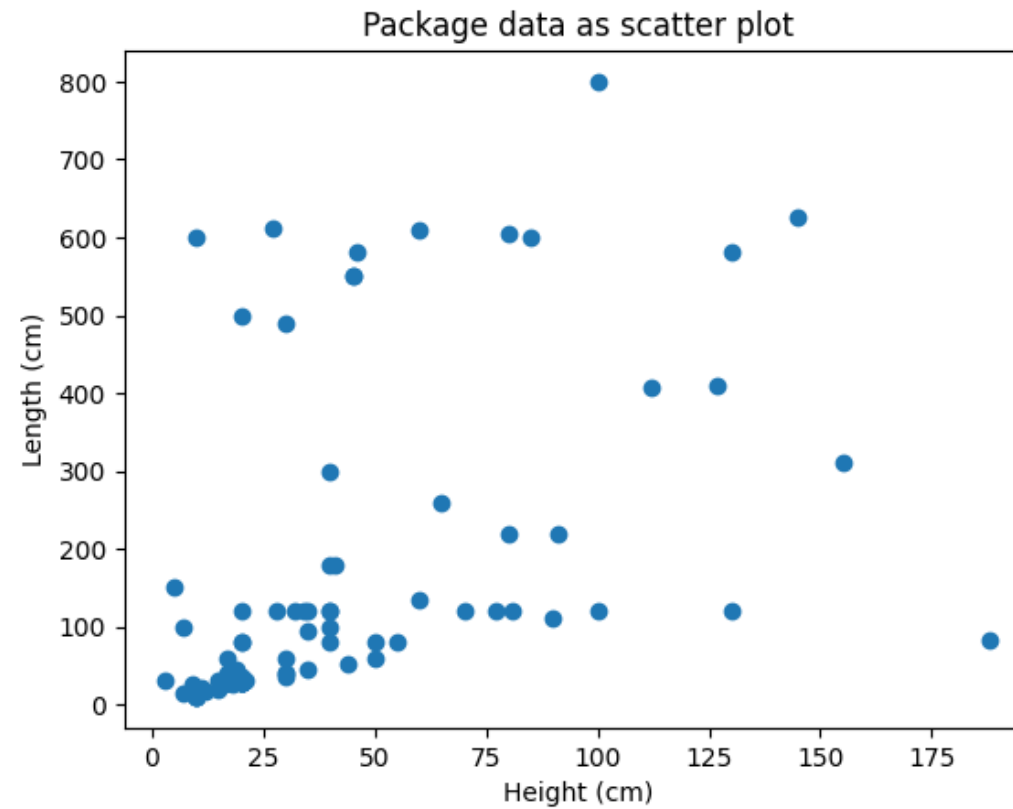


Data preprocessing tasks



Data reduction approaches

Daten visualisieren

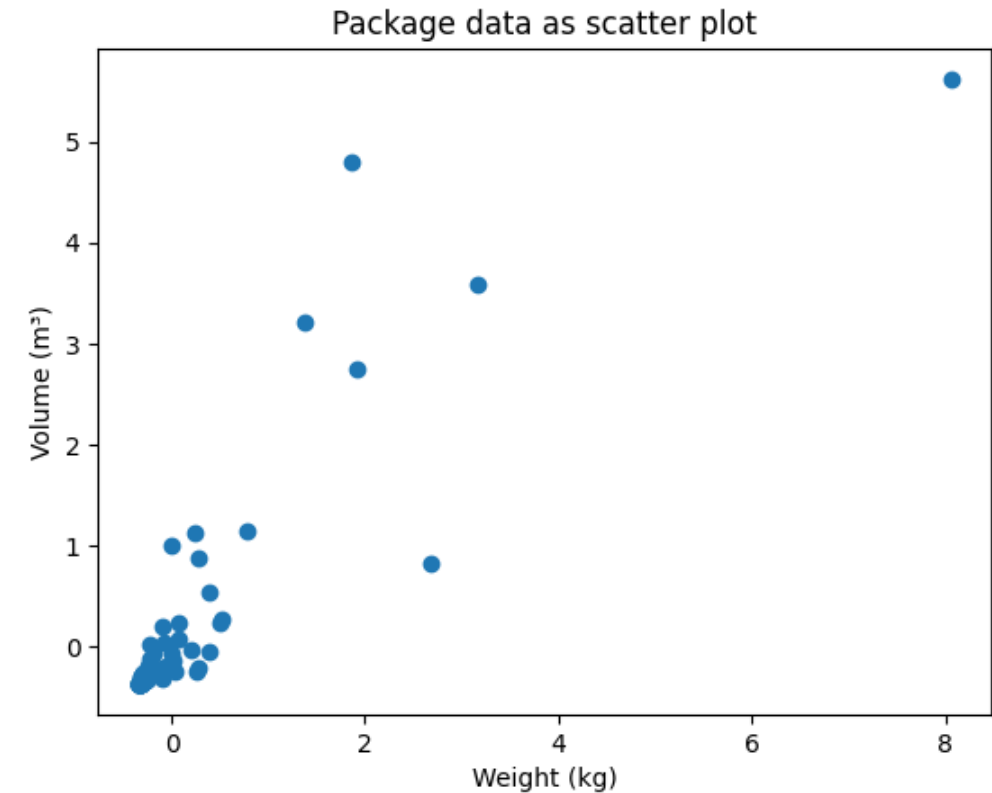
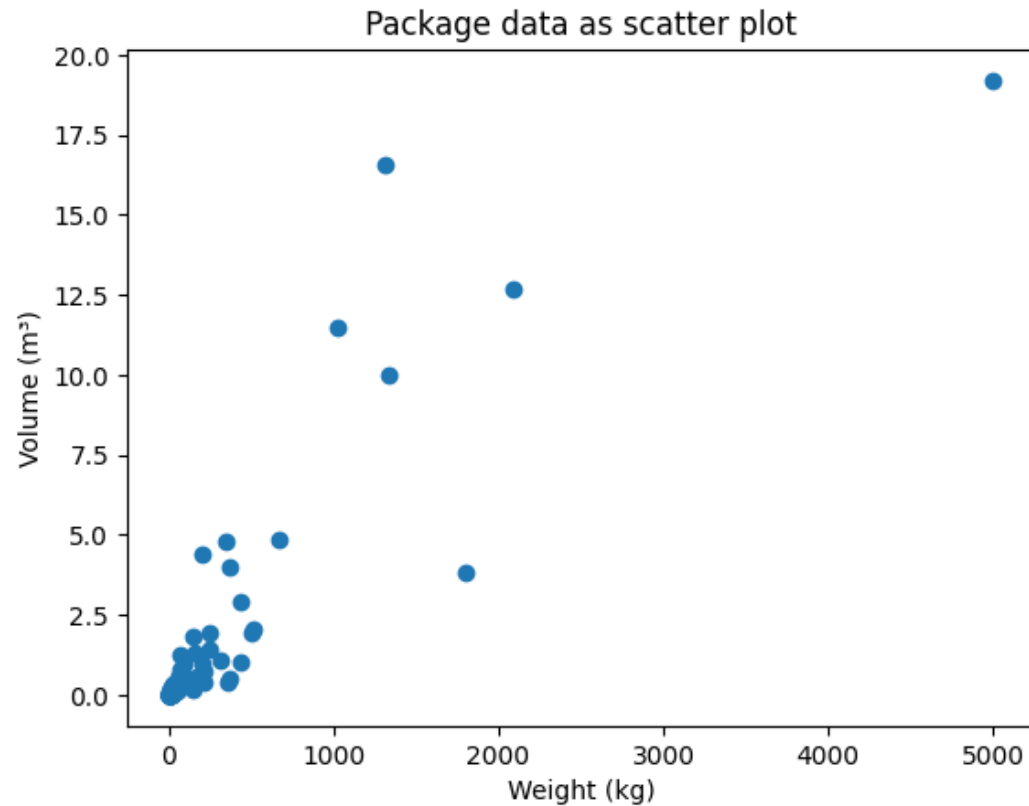


Visualisiert mit Matplotlib

Daten aufbereiten

- In diesem Datensatz
 - **Data Cleaning**, bspw. 1.001,57 zu 1001.57
 - **Data Integration**, z.B. Volumen ausrechnen
 - **Data Normalization**
 - **Feature Selection**

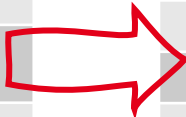
Daten visualisieren



Visualisiert mit Matplotlib

Daten aufbereiten

#	Package No	Shipment No	Gross Weight (kg)	Width (cm)	Height (cm)	Length (cm)
0	1007530-2011-03239	1000088	23	35	30	35
1	1007530-2011-03241	1000310	150	60	55	80
2	1007530-2011-03242	1000346	0,5	14	15	19
3	1007530-2011-03243	1000456	1,5	20	20	29
4	1007530-2011-03244	1000796	1	10	10	10
5	1007530-2011-03245	1000957	75	82	81	120
6	1007530-2011-03246	1000957	41	80	34	120
7	1007530-2011-03247	1001184	1.340	220	112	406
8	1007530-2011-03249	1001408	0,5	20	20	29
9	1007530-2011-03250	1001563	5	45	35	45



#	Gross Weight (kg)	Width (cm)	Height (cm)	Length (cm)	Volume (cm³)
0	23.0	35.0	30	35	36750.0
1	150.0	60.0	55	80	264000.0
2	0.5	14.0	15	19	3990.0
3	1.5	20.0	20	29	11600.0
4	1.0	10.0	10	10	1000.0
5	75.0	82.0	81	120	797040.0
6	41.0	80.0	34	120	326400.0
7	1340.0	220.0	112	406	10003840.0
8	0.5	20.0	20	29	11600.0
9	5.0	45.0	35	45	70875.0

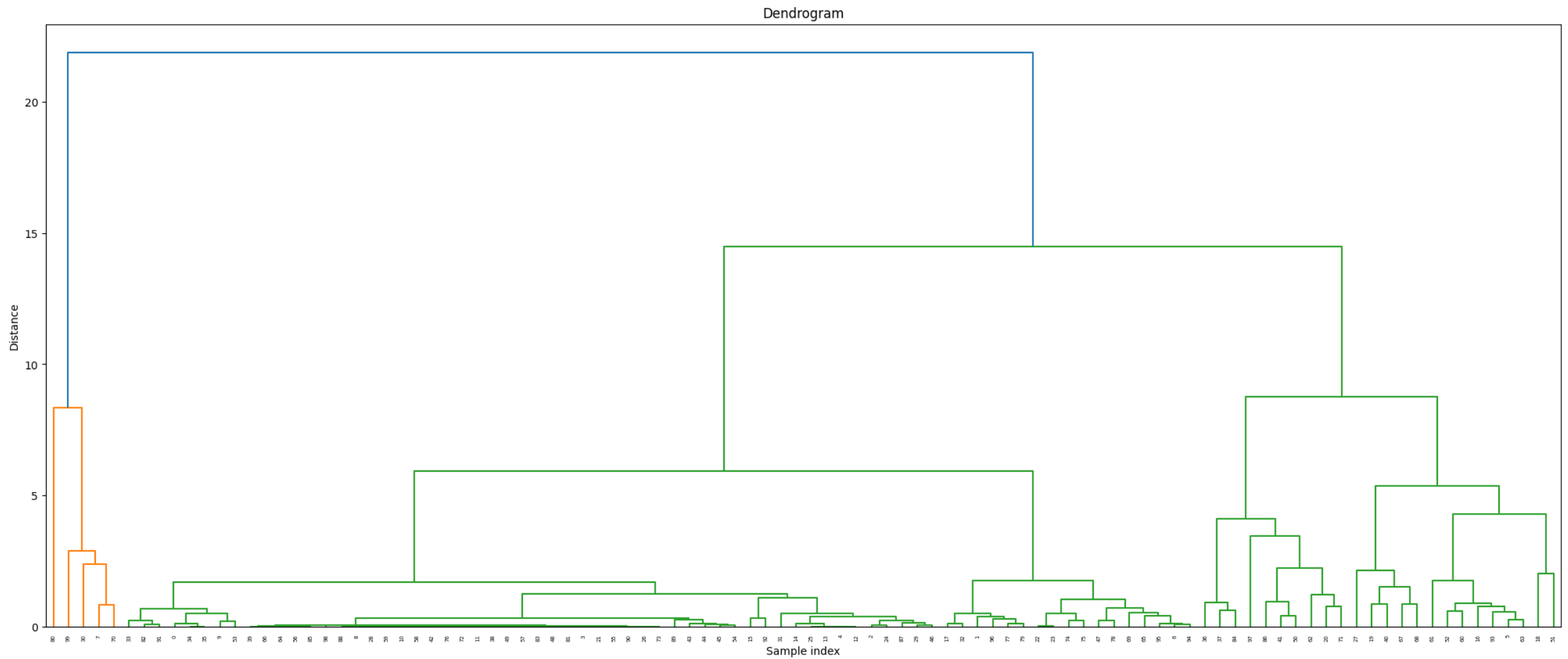
Original-Datensatz (erste zehn Zeilen)

Datensatz nach Aufbereitung (erste zehn Zeilen)

Cluster-Analyse: Hierarchisch

- Nach Schonlau, Matthias (2002) dient das **Dendrogramm** zur übersichtlichen Darstellung der **hierarchischen** Cluster-Bildung.
 - Wie Baumstruktur aufgebaut, vergleichbar mit Familienstammbaum

Cluster-Analyse: Hierarchisch



Visualisiert mit Matplotlib (Linkage ist 'ward')

Cluster-Analyse: Hierarchisch

- Sasirekha, K./Baby, P. (2013) beschreiben zwei unterschiedliche **Vorgehensweisen** zur Cluster-

Bildung:

- **Divisive**, d.h. von oben nach unten, d.h. von einem Cluster rekursiv nach unten aufteilen
- **Agglomerative**, d.h. von unten nach oben, d.h. jede Observierung bekommt zu Beginn ein eigenes Cluster und werden immer weiter verschmolzen

Cluster-Analyse: Hierarchisch

- Sasirekha, K./Baby, P. (2013) zählen folgende Verfahren auf, um die **Distanz** zwischen zwei

Observationen zu messen:

- **Euklidische Distanz**
- Quadratische euklidische Distanz (nicht in scikit-learn)
- **Manhattan Distanz**
- Maximum Distanz (nicht in scikit-learn)
- Mahalanobis Distanz (nicht in scikit-learn)
- **Kosinus Ähnlichkeit**

Cluster-Analyse: Hierarchisch

- Carvalho, Alexandre X. Y. u. a. (2009) beschreiben zwei weitere Distanz-Metriken:
 - **L2** (euklidische Norm)
 - **L1** (Summennorm)

Cluster-Analyse: Hierarchisch

- Murtagh, F. (1983) beschreibt mehrere Methoden, anhand welchen die **Cluster-Bildung** abhängig gemacht werden kann (engl. **Linkage**):
 - **Single linkage** (minimaler Abstand)
 - **Complete linkage** (maximaler Abstand)
 - **Average linkage** (Mittelwert)
 - **Median linkage**
 - **Centroid linkage** (Cluster-Schwerpunkte)
 - **Ward's linkage** (min. Zuwachs totaler Varianz)

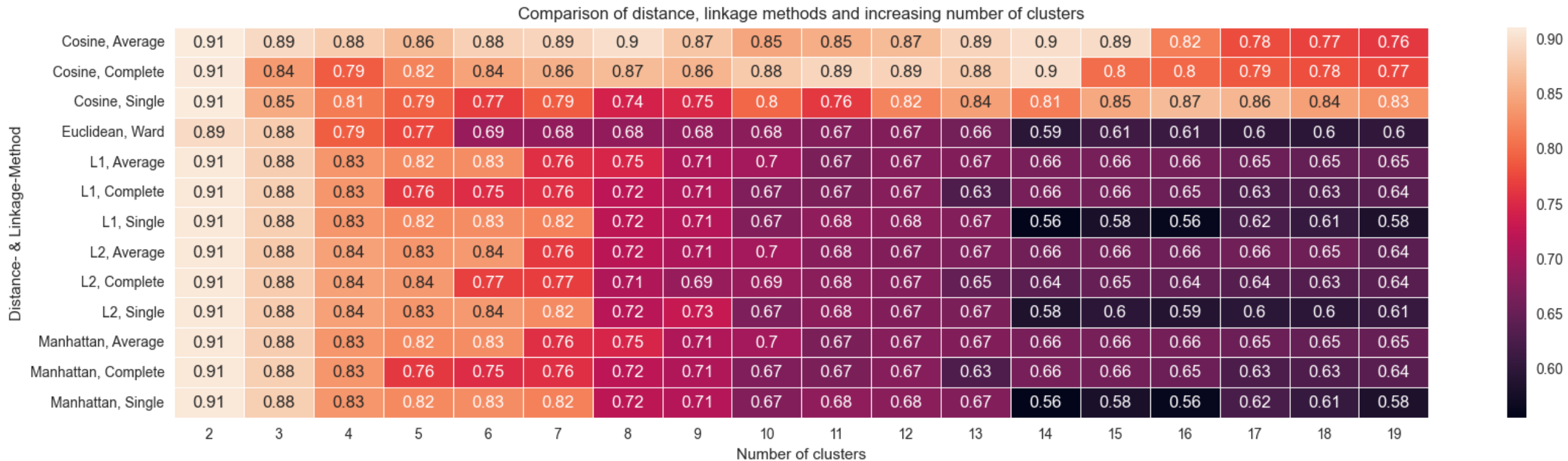
Cluster-Analyse: Hierarchisch

- Shahapure, Ketan R./Nicholas, Charles (2020) zeigen eine Metrik für die Bewertung eines

Clustering auf: der **Silhouette-Score**

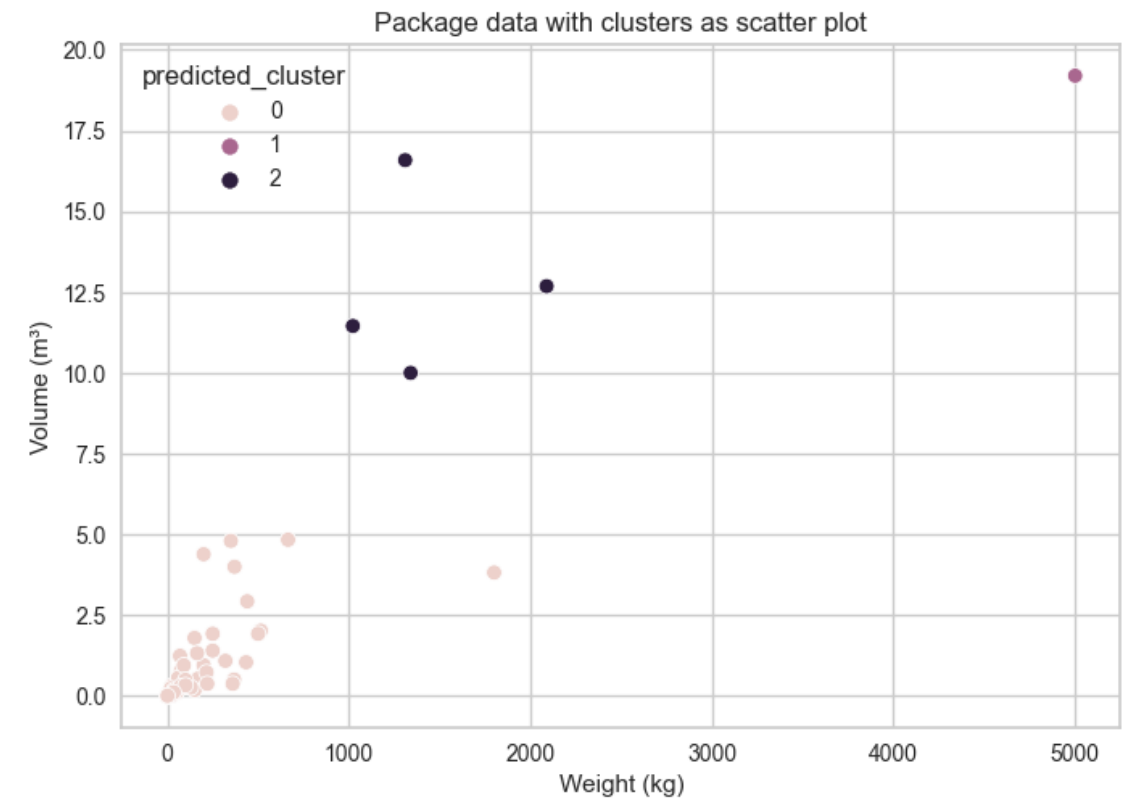
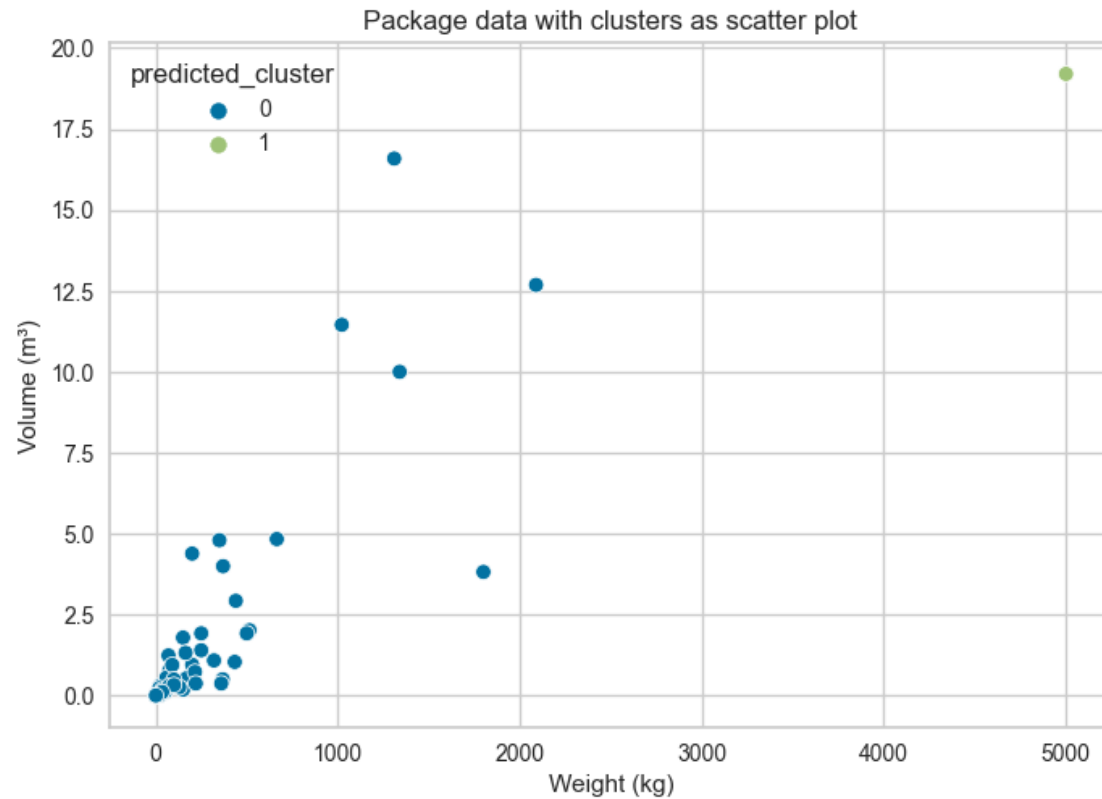
- Er ist der Mittelwert aller Silhouetten-Koeffizienten der Observationen
- Silhouette-Score nahe **1** => Daten sind in korrekten Clustern
- Silhouette-Score nahe **0** => mögliche Überlappung von Clustern
- Silhouette-Score nahe **-1** => Daten sind in falschen Clustern

Cluster-Analyse: Hierarchisch



Visualisiert mit Seaborn

Cluster-Analyse: Hierarchisch

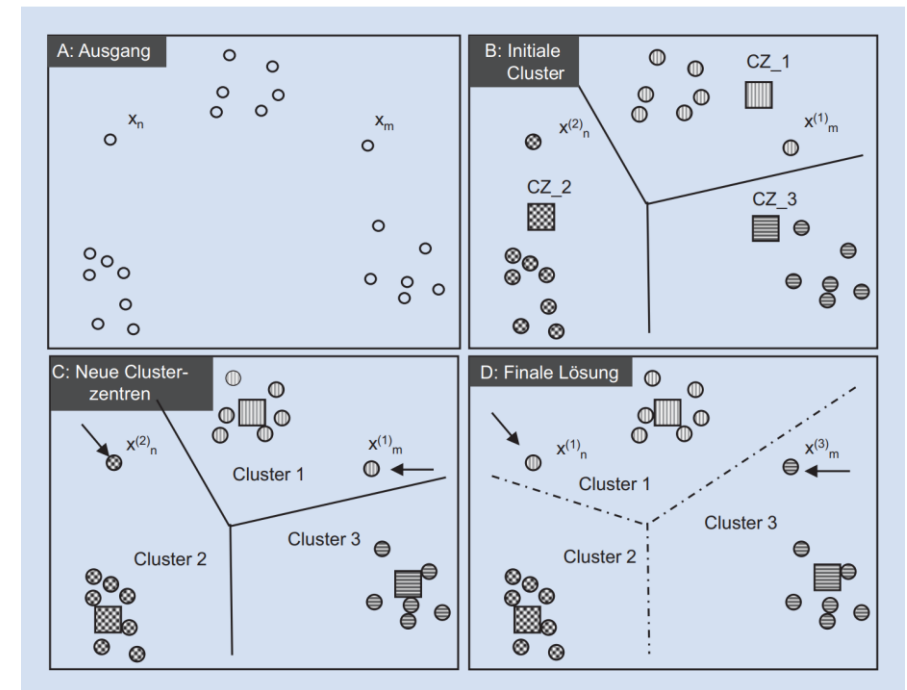


Visualisiert mit Matplotlib

Cluster-Analyse: k-Means

- **Partitionierung** eines Datensatzes in ***k* Cluster**
- **Zufällige** Definition von ***k* Clusterzentren**
- **Zuordnung der Datenpunkte** basierend auf der **euklidischen Distanz** $\sqrt{\sum_{i=1}^n (q_i - p_i)^2}$
- Neuberechnung der **Clusterzentren** als **Mittelwert** aller Datenpunkte innerhalb eines Clusters

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

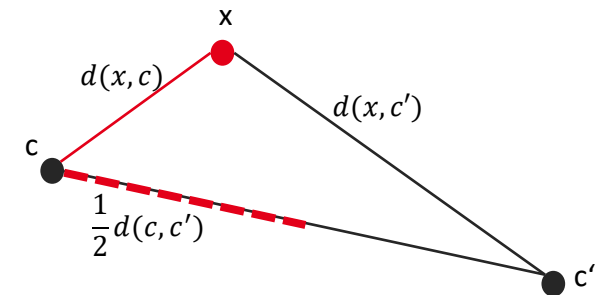


Backhaus, K (2021), S.567

Cluster-Analyse: k-Means

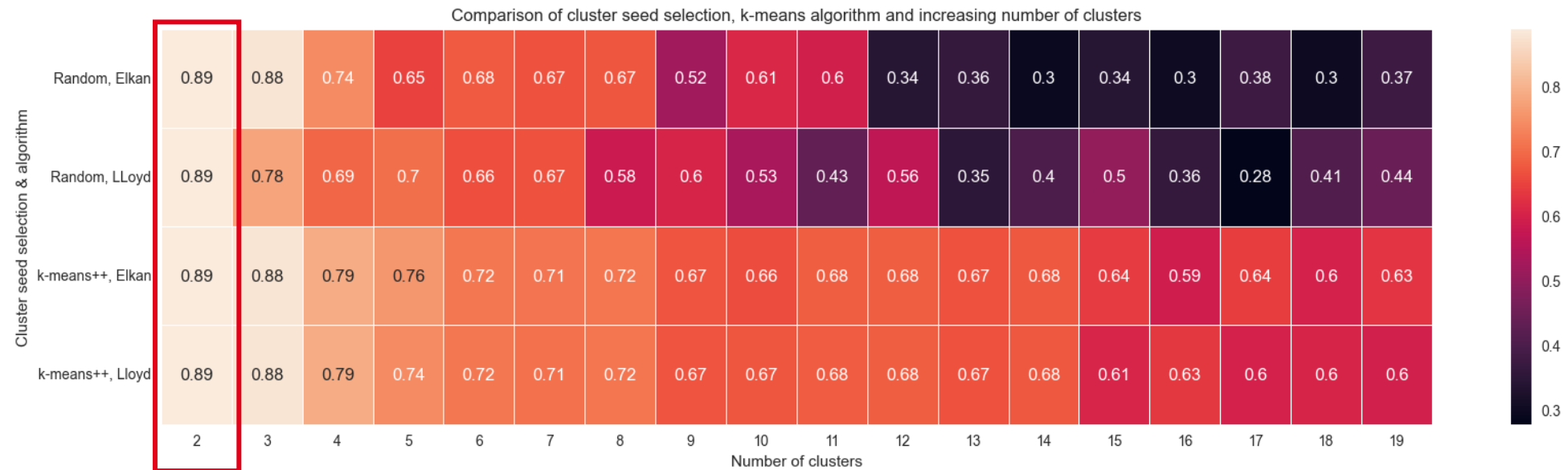
Optimierungen der k-means Methode

- Verbesserte Wahl der initialen Clusterzentren mit **k-means++** (Arthur, D./Vassilvitskii, S. (2007))
 - Auswahl des ersten Clusterzentrums mit gleichförmiger Wahrscheinlichkeit im gesamten Datenraum X
 - Auswahl der weiteren Clusterzentren nach einer Wahrscheinlichkeit von $\frac{D(x)^2}{\sum_{x \in X} D(x)^2}$
- Beschleunigung der Clusterzuordnung durch den **Elkan-Algorithmus** Elkan, C. (2003)
 - Reduzierung der Berechnungen durch Anwendung der Dreiecksungleichung
 - Wenn $\frac{1}{2} d(c, c') \geq d(x, c)$ dann $d(x, c') \geq d(x, c)$



➤ Verwendung des **Silhouette-Score** zur Identifikation der besten k-means Methode

Cluster-Analyse: k-Means

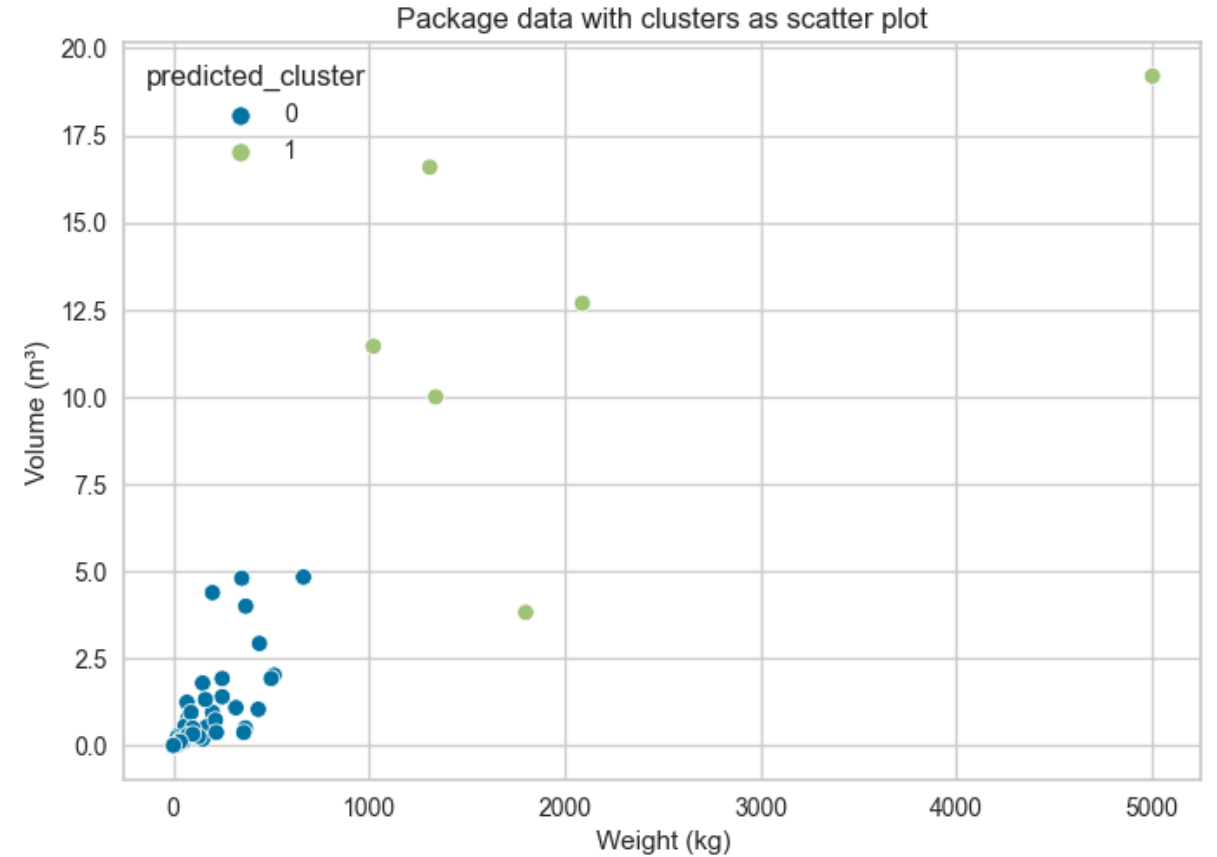


Visualisiert mit Seaborn

Cluster-Analyse: k-Means

Entsprechend des Ergebnisses der Heatmap:

- Verwendung von **k-means++** und **Elkan**
- Vorgabe von **$k = 2$** Clusterzentren



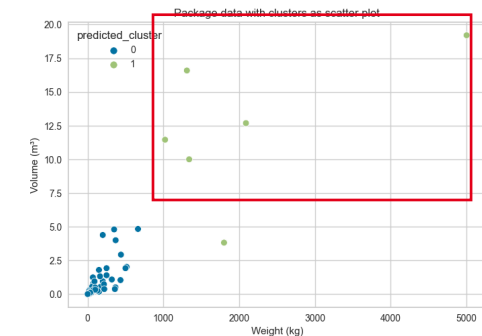
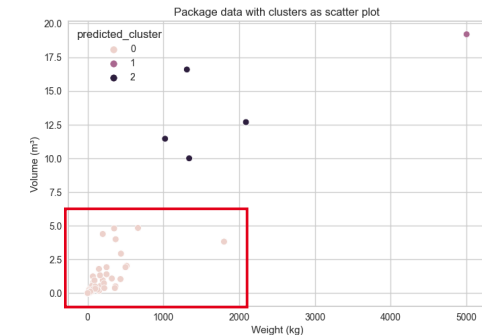
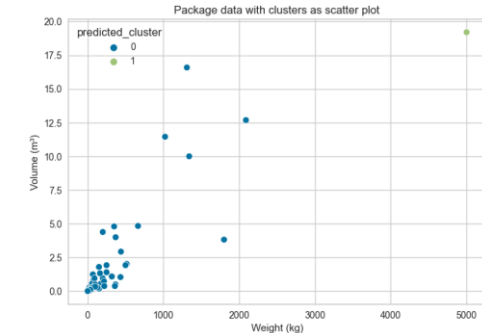
Visualisiert mit Matplotlib

Fazit

- Lessons-Learned #1: Zu verwendete Cluster-Methode hängt von **Verteilung** der Daten ab. Die Daten-**Aufbereitung** ist daher von zentraler Bedeutung
- Lessons-Learned #2: Der **Silhouette-Score** ist eine wichtige Metrik zur Bewertung des Clusterings
- Ausblick: Weitere mögliche Cluster-Methoden: GMM, DBSCAN, ...
- Relevant: Interpretation des Ergebnisses

Interpretation des Ergebnisses

- Welche Gruppen gleichartiger Packstücke können gebildet werden, um diese mit spezialisierten Teams zu bearbeiten?
- Unterschiedliche Ergebnisse der hierarchischen und k-means Methode
- Umsetzbarkeit in der Praxis zu berücksichtigen
 - Anzahl / Qualifikation der Mitarbeiter
 - Unterschiede in Bearbeitungsprozessen (z.B. Sondergenehmigungen)
 - Externe Einschränkungen (z.B. DHL Maximalgewicht 31,5 kg)
- Bilden von 2 Gruppen:
 - „Paket- und Speditionsware“
 - „Großware und Sondertransporte“



Quellen

Arthur, D.; Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia, PA, USA. pp. 1027–1035

Backhaus, K. et al., 2021, Multivariate Analysemethoden: Eine anwendungsorientierte Einführung. Wiesbaden: Springer Verlag, pp.491

Bacher, J., Pöge, A., Wenzig, K., 2010, Clusteranalyse: Anwendungsorientierte Einführung in Klassifikationsverfahren. München: Oldenbourg Wissenschaftsverlag GmbH

Carvalho, A.X.Y., Albuquerque, P.H.M., de Almeida Junior, G.R. and Guimaraes, R.D., 2009. Spatial hierarchical clustering. Revista Brasileira de Biometria, 27(3), pp.411-442.

Elkan, C. (2003), Using the Triangle Inequality to Accelerate k-means. In Proceedings of the Twentieth International Conference in Machine Learning (ICML-2003), Washington DC, 2003

García, S., Ramírez-Gallego, S., Luengo, J., Benítez, J.M. and Herrera, F., 2016. Big data preprocessing: methods and prospects. Big Data Analytics, 1(1), pp.1-22.

Murtagh, F., 1983. A survey of recent advances in hierarchical clustering algorithms. The computer journal, 26(4), pp.354-359.

Sasirekha, K. and Baby, P., 2013. Agglomerative hierarchical clustering algorithm-a. International Journal of Scientific and Research Publications, 83(3), p.83.

Schonlau, M., 2002. The clustergram: A graph for visualizing hierarchical and nonhierarchical cluster analyses. The Stata Journal, 2(4), pp.391-402.

Shahapure, K.R. and Nicholas, C., 2020, October. Cluster quality analysis using silhouette score. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 747-748). IEEE.

Shindler, M., no date, Approximation Algorithms for the Metric k-Median Problem, Zugriff via <http://www.cs.ucla.edu/~shindler/shindler-kMedian-survey.pdf>, 2022-11-27