Statistik-Übung 8 – *k*-means-Clustering

Beispielhafter Methoden- und Ergebnisteil

Methoden

Ziel war es, die 50 US-amerikanischen Bundesstaaten nach ihren Kriminalitätsprofilen zu gruppieren. Es lagen für alle Bundesstaaten Häufigkeiten von 7 Straftatbeständen vor (murder, rape, robbery, assault, burglary, theft, vehicle).

Die Analysen wurden in R, Version 4.2.2, durchgeführt (R Core Team 2024). Da sich die Raten um mehrere Zehnerpotenzen unterschieden (mittlere Mordrate: 6.8, mittlere Diebstahlrate: 2918 pro 100,000 Einwohner), wurden alle Daten zunächst *z*-transformiert. Die Gruppierung wurde mit *k*-means-Clustering durchgeführt und mit dem SSI-Kriterium die optimale Clusterzahl im Bereich zwischen 2 und 6 ermittelt (Befehl 'cascadeKM' im Package 'vegan'; Oksanen et al. 2022). Die Clusterzugehörigkeit wurde anschliessend in einer Hauptkomponentenanalyse des gleichen Datensatzes (Befehl 'fviz_cluster' im Package 'factoextra'; Kassambra & Mundt 2020) visualisiert. Abschliessend wurden die vier Cluster mittels Varianzanalyse mit HSD-Posthoc-Test bezüglich der mittleren Häufigkeit der sieben Kriminalitätsarten verglichen.

Ergebnisse

Die *k*-means-Clusterung ergab nach dem SSI-Kriterium die beste Lösung bei vier Clustern, die je zwischen 6 und 16 Bundesstaaten umfassten (Abb. 1). Auf der ersten Ordinationsebene waren die vier Cluster klar getrennt (Abb. 1). Die Varianzanalysen zeigten für alle sieben Kriminalitättypen höchstsignifikante Unterschiede zwischen den Clustern ($p < 0.001$). Nach den Posthoc-Tests (Abb. 2) waren nur im Fall von Raub alle Cluster paarweise voneinander verschieden, mit einer von Cluster 1 bis Cluster 4 ansteigenden Rate (Abb. 2). Cluster 1 hatte für alle Kriminalitätsarten die niedrigsten Raten, während die höchsten Raten in fünf Fällen im Cluster 3 und in drei Fällen in Cluster 4 auftraten (Abb. 2).

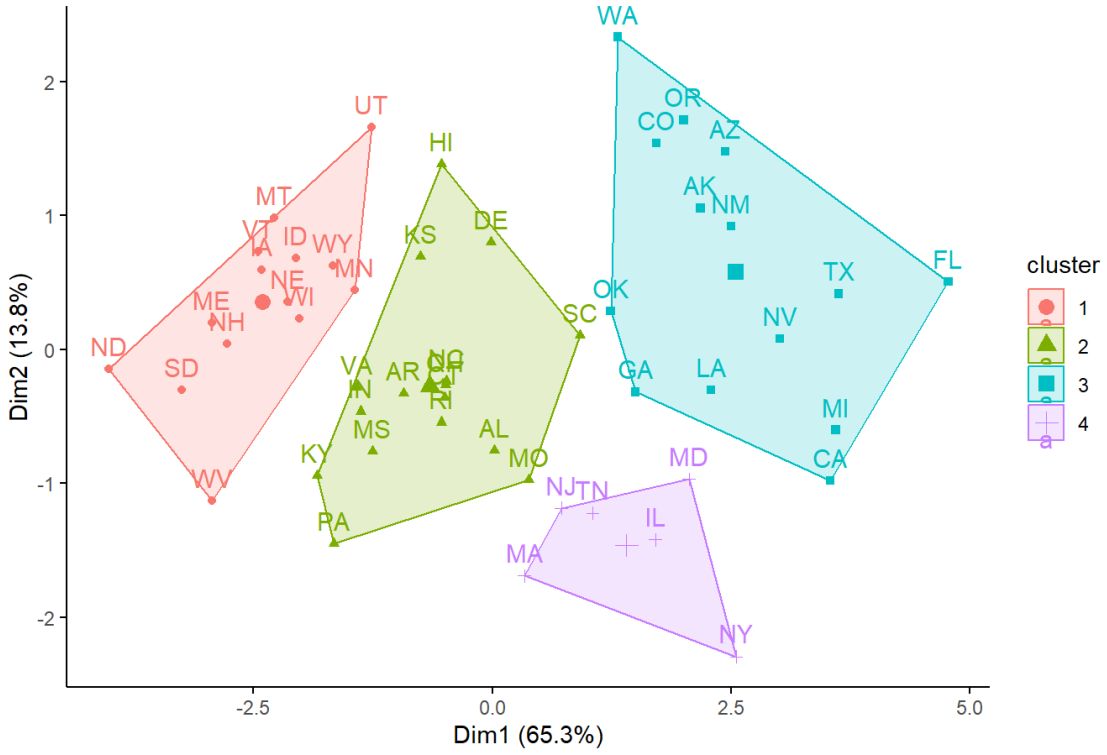

Abb. 1. Visualisierung der 4-Clusterlösung mit *k*-means-Clustering der Kriminalitätsprofile der 50 US-amerikanischen Bundesstaaten. Die vier Cluster und die zugehörigen Bundesstaaten sind auf der ersten Ordinationsebene der Hauptkomponentenanalyse der *z*-transformierten Werte dargestellt.

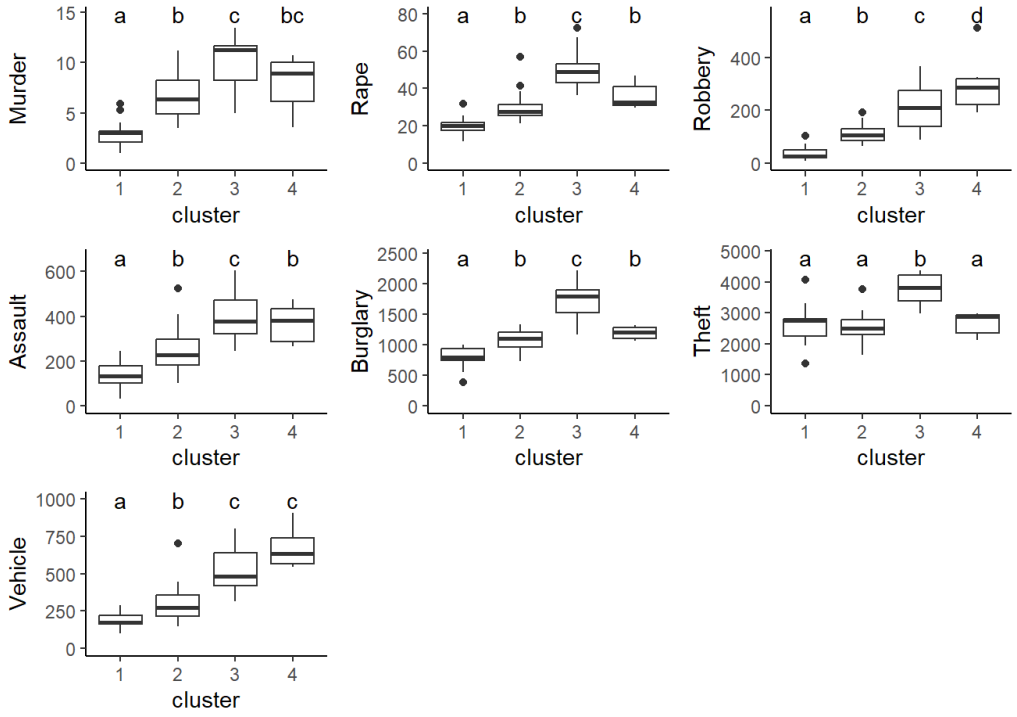

Abb. 2. Boxplots und post-hoc-Tests der Häufigkeit von Kriminalitätsraten im Vergleich der vier unterschiedenen Cluster von Bundesstaaten

Quellen

Kassambra, A. & Mundt, F. (2020) *factoextra: extract and visualize the results of multivariate data analyses. Version 1.0.7*. URL: https://CRAN.R-project.org/package=factoextra.

Oksanen, J., Simpson, G., F.G., Kindt, R.,Legendre, P., Minchin, P.R., O'Hara, R.B. et al. (2022) *vegan: community ecology package*. Version 2.6-4. URL: https://CRAN.R-project.org/package=vegan.

R Core Team. (2024) *R: A language and environment for statistical computing. R Foundation for Statistical Computing*. http://www.r-project.org/.