R-Tutorium
Übung 7

Wintersemester 2020/21
Christoph Herrmann

Aufgabe 1:
Laden Sie den Datensatz *biathlon3* aus der ILIAS Übungseinheit und alle weiteren Pakete, die Sie für die Bearbeitung der Aufgaben benötigen. Ziel der Aufgabe ist es, den gegebenen Datensatz deskriptiv zu analysieren und anschließend eine hierarchische und partitionierende Clusteranalyse durchzuführen. Der folgende Datensatz beschreibt eine Zusammenstellung von Biathlon- Rennergebnissen der Saison 2017/18, dem Renntyp-abhängigen Streckendistanzen und der Anzahl an abgegebenen Schüssen. Die Daten stammen von der Internetseite der Internationalen Biathlon Union (`https://www.biathlonworld.com/`). Biathlon ist ein vielseitiger Sport und sehr beliebt. Er kombiniert zwei Disziplinen (Langlauf und Schießen). Tabelle 1 zeigt die Variablen des Datensatzes. Dieser besteht aus 3612 Beobachtungen und 21 Variablen.

(a) Betrachten Sie zunächst alle Variablen im Datensatz und analysieren Sie diese hinsichtlich Ihrer Lage und Streuung. Lassen sich irgendwelche Auffälligkeiten feststellen? Wenn ja sollten Sie überlegen, wie Sie diese bereinigen bzw. beseitigen könnten.

(b) Stellen Sie als nächstes die Variablen eindimensional graphisch dar. Nutzen Sie hierfür die unterschiedlichen Funktionen, die Sie in den Tutorials rund um den *tidyverse*-Approach kennengelernt haben.

(c) Überlegen Sie im Anschluss, welche Merkmale für eine Clusteranalyse geeignet sind (metrisches Skalenniveau). Erzeugen Sie in einem nächsten Schritt einen Datensatz, der nur die zur Analyse genutzten Variablen enthält und standardisieren Sie die Daten. Überlegen Sie vor dem Standardisieren, ob es Sinn macht, die Streckendistanzen mit einzubeziehen und gegebenenfalls die Streckenzeiten zu transformieren.

(d) Führen Sie in einem nächsten Schritt eine hierarchische Clusteranalyse auf Basis der euklidischen Distanz und des single-linkage, sowie Ward Verfahrens durch. Stellen Sie hier wieder Auffälligkeiten fest? Wenn ja überlegen Sie auch hier, wie Sie mit den auffälligen Beobachtungen umgehen und führen Sie die Analyse gegebenenfalls erneut durch. Entscheiden Sie sich nun für eine Anzahl an Clustern, die Ihnen angemessen erscheint.

(e) Führen Sie nun eine Clusteranalyse auf Basis eines partitionierenden Verfahrens durch und entscheiden Sie sich auch hier für eine angemessene Zahl an Clustern.

(f) Fügen Sie anschließend die Zuordnung der Cluster und die vor der Analyse entfernten Variablen dem Datensatz bei. Lassen sich in den Clustern gewisse Strukturen wiederfinden? Nutzen Sie hierfür deskriptive Methoden und stellen Sie Ihre Ergebnisse graphisch dar.

(g) Überlegen Sie, ob es sinnvoll sein könnte, die Daten vor der Analyse in ein männliches und weibliches Subset einzuteilen.

Merkmal	Beschreibung
gender	beschreibt das Geschlecht des Biathleten und besitzt die Ausprägungen W für Woman und M für Man.
competition	beschreibt den Renntyp und hat die Ausprägungen I für Individual, S für Sprint, M für Mass Start und P für Pursuit.
type	beschreibt den Wettkampftyp und hat die Ausprägungen W für World Cup, C für Championship und O für Olympic Game.
nation	beschreibt die nationale Zugehörigkeit der Athleten und enthält die englischsprachigen Länderkürzel.
total.time	beschreibt die benötigte Zeit in Sekunden bis zum Ziel.
course.lap.1, course.lap.2, course.lap.3, course.lap.4, course.lap.5 und course.total	beschreiben für die jeweiligen Runden die benötigten Zeiten in Sekunden, die alleine für die Langlaufdisziplin benötigt wurden. Schießzeiten und der Weg vom Eingang bis zum Ende der Schießanlage sind nicht beinhaltet.
shoot.times.1, shoot.times.2, shoot.times.3, shoot.times.4 und shoot.times.total	beschreiben für die jeweiligen Schießdisziplinen die benötigten Zeiten in Sekunden. Laufzeiten und der Weg vom Eingang bis zum Ende der Schießanlage sind nicht beinhaltet.
fails.1, fails.2, fails.3, fails.4 und fails.total	beschreiben die Anzahl derFehlschüsse bei den jeweiligen Schießdisziplinen.

Tabelle 1: Variablenbeschreibung des Datensatzes