Programmieren mit Statistischer Software **5. Block**
Grün **SS 2017**

5 Ein- und Ausgabe von Daten

Aufgabe 1:

Laden Sie den Wholesale customers Datensatz, der im UCI Machine Learning Repository unter http://archive.ics.uci.edu/ml/datasets/Wholesale+customers zur Verfügung steht, herunter und lesen Sie ihn nach R ein.

- Wieviele Beobachtungen und Variablen enthält der Datensatz?
- Kodieren Sie die Variablen „Channel“ und „Region“ als nominale Variable mit sprechenden Namen für die Merkmalsausprägungen.
- Extrahieren Sie die Teilmenge der Beobachtungen, wo die Region „Lisbon“ ist und Grocery Ausgaben über dem Median sind.

Aufgabe 2:

Laden Sie den Labor Relations Datensatz, der im UCI Machine Learning Repository unter http://archive.ics.uci.edu/ml/datasets/Labor+Relations zur Verfügung steht, herunter und lesen Sie ihn nach R ein. Nehmen Sie die Datendatei `labor-neg.data`, die im Verzeichnis C4.5 zur Verfügung gestellt wird.

- Achten Sie darauf, dass Sie die fehlenden Werte richtig einlesen.
- Kodieren Sie die Variablen Urlaub (“vacation”), Beitrag zur Zahnversicherung (“contribution to dental plan”) und Beitrag zur Krankenversicherung (“contribution to health plan”) richtig als ordinale Variablen.

Hinweis: Geordnete Faktoren bekommt man, indem man bei `factor(..., ordered = TRUE)` verwendet.

Aufgabe 3:

Verwenden Sie den Datensatz Labor Relations, um folgende Analysen zu machen:

- Vergleichen Sie, wie hoch der Anteil an fehlenden Werten bei den Variablen je nach Abschluss (“settlement”) ist.
- Bestimmen Sie für jene Variablen, die in keiner der beiden Abschlussgruppen mehr als 50% fehlende Werte haben und die metrisch sind, die Mittelwerte für die beiden Gruppen.

Aufgabe 4:

Auf der Seite der Statistik Austria stehen Daten, die von der Statistik Austria erhoben wurden, frei zur Verfügung. Wir wollen uns den Datensatz über ie Bevölkerungsentwicklung seit 1869 anschauen. Diese finden Sie unter

https://www.statistik.at/web_de/statistiken/menschen_und_gesellschaft/bevoelkerung/bevoelkerungsstruktur/bevoelkerung_nach_alter_geschlecht/index.html

beim Punkt „Ergebnisse im Überblick: Bevölkerung seit 1869“ als Datei `ergebnisse_im_ueberblick_bevoelkerung_seit_1869.xlsx`.

- Lesen Sie den Datensatz nach R ein, um die folgenden 3 Datensätze zu erzeugen:

1. Bevölkerung pro Jahr zu Jahresbeginn seit 2002.
2. Bevölkerung pro Jahr und Geschlecht zu Jahresbeginn seit 2002, d.h. mit 3 Spalten: Jahr, Geschlecht, Bevölkerung.
3. Bevölkerung pro Jahr und Alter zu Jahresbeginn seit 2002, d.h. mit 3 Spalten: Jahr, Alter, Bevölkerung.

Dafür ist es wahrscheinlich am einfachsten, wenn Sie die Daten erst in einem Tabellenkalkulationsprogramm, das `.xlsx` Dateien öffnen kann, bearbeiten, sodass Sie die Daten dann leicht einlesen können.

- Überprüfen Sie, ob die Gesamtanzahl sowie die Gesamtanzahl pro Jahr jeweils übereinstimmt.
- Bestimmen Sie die Anteile der unter 14-jährigen und über 65-jährigen an der Gesamtbevölkerung über die Zeit und interpretieren Sie sie.

Aufgabe 5:

Im Folgenden verwenden wir die Werte der Volkszählungen aus dem Datensatz von der Homepage der Statistik Austria.

Die zeitlichen Abstände zwischen den Volkszählungen sind unregelmässig. Daher ist ein direkter Vergleich der prozentuellen Änderungen schwer interpretierbar. Jedoch kann aus den Stichtagen die Länge der Zeitspanne bestimmt werden, wo die Änderungen eingetreten sind.

- Auf der Wikipedia-Seite (https://de.wikipedia.org/wiki/Liste_der_Volksz%C3%A4hlungen_in_%C3%96sterreich) finden sich die Stichtage der jeweiligen Volkszählung. Erzeugen Sie also erst einmal einen Datensatz mit den Variablen Jahr und Stichtag.
- Verbinden Sie die Information mit den Volkszähldaten und bestimmen Sie die prozentuellen Änderungen zwischen den Volkszählungen.
- Bestimmen Sie damit zuerst die tägliche Wachstumsrate pro Zeitspanne und daraus durchschnittliche jährliche bzw. die 10-jährige Wachstumrate innerhalb jeder Periode und vergleichen Sie diese mit den prozentuellen Änderungen.
- In welcher Periode war das größte und in welcher das geringste Wachstum in Österreich?