

Übungsblatt 2

Datenanalyse und -visualisierung mit R
Hochschule Bonn-Rhein-Sieg

Pekka Sagner M.Sc.

22. Oktober 2021

Einführung in das Tidyverse

Aufgabe 1: »Ordnung muss sein«

- a) Neben Vektoren kennt R weitere (interne) Datenformate. Wir kommen im Laufe der Vorlesung auf die meisten zu sprechen. In der Regel arbeiten wir jedoch mit Datensätzen. In der modernen Coding-Philosophie, dem **Tidyverse**, sollten Daten einer bestimmten Struktur folgen - sie sollen **tidy** (ordentlich/sauber) sein. Wann ist ein Datensatz **tidy**? Lesen Sie zur Beantwortung der Frage [R4DS, Kapitel 12.1-12.2 \(Wickham/Grolemund, 2021\)](#) oder werfen Sie einen Blick in die Vorlesungsunterlagen.

Ein Datensatz ist **tidy**, wenn:

- 1)
 - 2)
 - 3)
- b) Wir werden ab jetzt im Kurs **immer** mit dem **Tidyverse** arbeiten, der Sammlung von Paketen, die die Arbeit mit **tidy**-Daten besonders einfach macht und speziell darauf zugeschnitten ist. Installieren Sie das Paket **tidyverse** mit dem Befehl `install.packages("tidyverse")`.
- c) Um die Inhalte eines Pakets, wie zum Beispiel Funktionen und Datensätze, für die Analyse zu nutzen, laden Sie dieses mit dem Befehl `library()`. Laden Sie das Paket **tidverse**.
- d) Das Paket **tidyverse** enthält Übungsdatensätze, auf die wir zugreifen können, nachdem wir das Paket geladen haben. Mit den Befehlen `table1`, `table2` und `table3` können Sie diese in der Konsole betrachten. Weisen Sie den drei Datensätzen Namen zu, z.B. **datensatz_1**, **datensatz_2**, **datensatz_3**. Betrachten Sie die Datensätze im **Viewer**.
- e) Beschreiben Sie den Aufbau der drei Datensätze in Worten.

Die Datensätze finden Sie hier auch noch einmal abgedruckt:

Datensatz 1:

| country | year | cases | population |
|-------------|------|-------|------------|
| Afghanistan | 1999 | 745 | 19987071 |
| Afghanistan | 2000 | 2666 | 20595360 |

| country | year | cases | population |
|---------|------|--------|------------|
| Brazil | 1999 | 37737 | 172006362 |
| Brazil | 2000 | 80488 | 174504898 |
| China | 1999 | 212258 | 1272915272 |
| China | 2000 | 213766 | 1280428583 |

Datensatz 2:

| country | year | type | count |
|-------------|------|------------|------------|
| Afghanistan | 1999 | cases | 745 |
| Afghanistan | 1999 | population | 19987071 |
| Afghanistan | 2000 | cases | 2666 |
| Afghanistan | 2000 | population | 20595360 |
| Brazil | 1999 | cases | 37737 |
| Brazil | 1999 | population | 172006362 |
| Brazil | 2000 | cases | 80488 |
| Brazil | 2000 | population | 174504898 |
| China | 1999 | cases | 212258 |
| China | 1999 | population | 1272915272 |
| China | 2000 | cases | 213766 |
| China | 2000 | population | 1280428583 |

Datensatz 3:

| country | year | rate |
|-------------|------|-------------------|
| Afghanistan | 1999 | 745/19987071 |
| Afghanistan | 2000 | 2666/20595360 |
| Brazil | 1999 | 37737/172006362 |
| Brazil | 2000 | 80488/174504898 |
| China | 1999 | 212258/1272915272 |
| China | 2000 | 213766/1280428583 |

f) Welche der drei Datensätze sind nicht **tidy**? Warum nicht?

Aufgabe 2: Die Pipe - %>%

Pipes sind ein mächtiges Instrument, um eine Reihe von Operationen durchzuführen. Sie sind mit der Grund, warum R und das Tidyverse sich einer so großen Beliebtheit erfreuen. Das Ziel und der Grund für die Implementierung der Pipe ist es, Code besser schreib- und lesbar zu machen. Wir lesen %>% als »und dann«. Aus der klassischen Funktionsschreibweise $f(x)$, wird dadurch zum Beispiel `x %>% f()`. (Leseempfehlung: [R4DS, Kapitel 19 \(Wickham/Grolemund, 2021\)](#))

Stellen Sie sich vor, ein Bekannter beschreibt Ihnen seinen Tagesablauf wie folgt: *»Ich bin heute um 6:00 Uhr aufgestanden. Dann habe ich mir eine Hose und einen Pullover angezogen. Dann habe ich einen starken Kaffee getrunken. Danach bin ich nicht mit dem Auto, sondern mit dem Fahrrad zur Hochschule gefahren. Dann habe ich mit herausragender Laune gelernt.«*

- Schreiben Sie den beschriebenen Ablauf in einer klassischen funktionalen Form $f(g(h(...(x))))$ auf.
- Schreiben Sie die vorherige Szene um, nutzen Sie dafür die Pipe (%>%).

Aufgabe 3: Olympische Spiele

Wir sind bereit, mit unserem ersten »echten« Datensatz zu arbeiten. Der Datensatz enthält Informationen zu den modernen Olympischen Spielen, beginnend mit Athen (1896) bis Rio de Janeiro (2016) (Quelle: [tidytuesday](#)).

Zum Datensatz liegen uns folgende Variablenbeschreibungen vor:

| variable | class | description |
|----------|-----------|------------------------------------|
| id | double | Athlete ID |
| name | character | Athlete Name |
| sex | character | Athlete Sex |
| age | double | Athlete Age |
| height | double | Athlete Height in cm |
| weight | double | Athlete weight in kg |
| team | character | Country/Team competing for |
| noc | character | noc region |
| games | character | Olympic games name |
| year | double | Year of olympics |
| season | character | Season either winter or summer |
| city | character | City of Olympic host |
| sport | character | Sport |
| event | character | Specific event |
| medal | character | Medal (Gold, Silver, Bronze or NA) |

- Laden Sie den Datensatz **olympics.csv** aus dem LEA-Kursordner herunter. Speichern Sie diesen in einem Unterordner (z. B. *data*) des R-Projektordners.
- Laden Sie den Datensatz in Ihr R-Environment ein. Nutzen Sie hierfür die Funktion `read_csv()`. Betrachten Sie den Datensatz im Viewer.
- Wie viele Beobachtungen enthält der Datensatz? Was steckt in diesem Fall hinter einer Beobachtung?
- Wie viele Beobachtungen liegen für die Olympischen Sommerspiele 2016 vor?
- Erstellen Sie einen neuen Datensatz, der nur Beobachtungen zu den Olympischen Sommerspielen 2016 und nur die Variablen **id**, **name**, **sex**, **age**, **height**, **weight**, **team**, **sport**, **event** und **medal** enthält.
- Wir sind im Folgenden an Informationen zur Größe der AthletInnen interessiert.
 - Wie groß war der/die größte bzw. der/die kleinste AthletIn?
 - Wie viel größer ist der/die größte AthletIn als der/die kleinste AthletIn?
 - An welchem Wettkampf hat der/die größte bzw. kleinste AthletIn teilgenommen?
 - Wie groß waren die AthletInnen bei den Olympischen Sommerspielen 2016 im Durchschnitt?
 - Wie groß sind die AthletInnen der verschiedenen Sportarten im Durchschnitt? In welcher Sportart sind die AthletInnen durchschnittlich am kleinsten, in welcher am größten?
 - Wie groß sind die Athletinnen im Durchschnitt, wie groß sind die Athleten?
 - Wie groß sind die Athletinnen je Sportart im Durchschnitt und wie groß die Athleten?
 - Wie viele Athleten sind größer als 1,80 m? Wie viele Athletinnen sind größer als 1,60 m? Bestimmen Sie jeweils den Anteil in der Gruppe.
- Wie groß ist der durchschnittliche Body-Mass-Index (BMI) von weiblichen Volleyballspielerinnen im Alter zwischen 20 und 30 Jahren?