Einlesen von Datensätzen in unterschiedlichen Formaten

Cornelius Puschmann

Table of contents

Um in R mit größeren Datensätzen arbeiten zu können, müssen diese zunächst eingelesen werden. Um Daten in R einzulesen (d.h. um sie in ein R-Objekt im Arbeitsspeicher ihres Rechners zu überführen), gibt es verschiedene Möglichkeiten. Die gebräuchlichste Methode ist die Verwendung von Funktionen. Zum Glück bietet R eine Reihe nützlicher Funktionen für ganz unterschiedliche Dateiformate, z. B. CSV-, Excel-, SPSS- und STATA-Dateien.

1 Data Management

Zunächst laden wir die Pakete, die für den Import von Dateien in den o.g. Formaten nützlich sind.

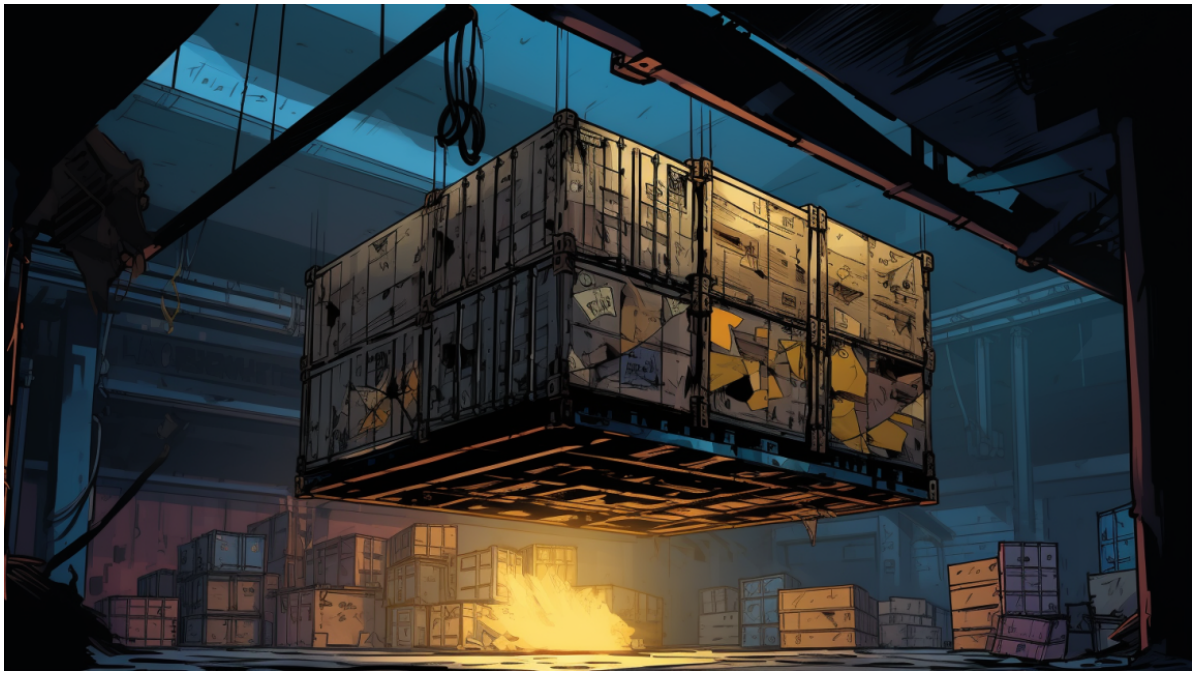

Figure 1: Wie bekomme ich meine Daten?, Bild generiert von Midjourney

```
if(!require("pacman")) {install.packages("pacman");library(pacman)}
p_load(readr, readxl, haven)

options(scipen = 999)
```

2 CSV-Dateien

CSV-Dateien sind eine sehr einfache Art von Daten, die in R eingelesen werden können. Sie sind als Textdateien gespeichert, wobei die Tabellenspalten durch Kommas (oder manchmal Semikolons oder Tab-Zeichen) getrennt sind. Um eine CSV-Datei in R einzulesen, können Sie die Funktion read_csv (bzw. hier read_csv2) aus dem Paket readr verwenden.

Dieser Code importiert die Datei geschlechterverteilung.csv in das R-Environment. Die Datei geschlechterverteilung.csv muss im gleichen Verzeichnis wie der R-Code gespeichert sein.

```
geschlechterverteilung_csv <- read_csv2("geschlechterverteilung.csv")
```

Die Funktion `read_csv2` liefert hier eine Menge zusätzlicher Informationen, etwa dazu, welche Datentypen für die eingelesenen Variablen gewählt wurden. Es wurde hier deshalb nicht

`read_csv` sondern `read_csv2` gewählt, weil letztere das Semikolon (;) als Trennzeichen und das Komma (,) als Dezimalzeichen verwendet. Hingegen geht read_csv vom Komma als Trennzeichen und dem Punkt als Dezimalzeichen aus, wie im angelsächsischem Gebrauch üblich. Liest man also mit der falschen Funktion ein, sind die Daten unter Umständen fehlerhaft.

3 Excel-Dateien

Ebenfalls nützlich ist die Möglichkeit, Daten aus Microsoft Excel in R zu imoportiern. Excel-Dateien können mit der Funktion `readxl::read_excel()` in R eingelesen werden. Diese Funktion unterstützt alle gängigen Excel-Formate, einschließlich XLSX, XLS und CSV.

```
geschlechterverteilung_excel <- read_excel("geschlechterverteilung.xlsx")
```

4 SPSS-Dateien

SPSS ist eine statistische Softwareanwendung, die von er Firma IBM (früher SPSS Inc) für Datenmanagement, inferenzstatistische Analysen und (kommerziell) Business Intelligence entwickelt wurde.SPSS-Dateien können mit der Funktion `haven::read.spss()` in R eingelesen werden. Diese Funktion unterstützt alle gängigen SPSS-Formate, einschließlich SAV und DSP.

Da der ALLBUS im konkreten Fall im Stata-Format verwendet wird, verwenden wir für SPSS ein anderes Beispiel, nämlich die dritte Welle (in 2016 erhoben) des Global Report on Adult Learning and Education (GALE-3) der UNESCO.

```
gale3_spss <- read_spss("https://uil.unesco.org/i/doc/adult-education/grale-3/survey-data/gra
```

Das Beispiel zeigts dass wir neben lokal abgespeicherten Datein auch problemlos Web-Adressen als Pfadangabe beim Import von Daten verwenden können, wenn diese direkt auf die relevanten Daten zeigen.

5 Stata-Dateien

Bei STATA handelt es sich um eine weitere kommerzielle Statistik-Software. Mittels Stata analysiert man einfache und komplexe Datenmodelle. Die Software gehört neben SPSS zu den bekanntesten Programmen für die professionelle Datenauswertung. STATA-Dateien können mit der Funktion `haven::read_dta()` in R eingelesen werden. Diese Funktion unterstützt alle gängigen STATA-Formate, einschließlich DTA und DTB.

Wir verwnden als Beispiel für den Import von STATA-Daten hier den ALLBUS, da dieser im STATA-Format vorliegt.

```
allbus_stata <- read_dta("Datensatz/Allbus_2021.dta")
```

6 Andere (textbasierte) Formate

Für andere Datenformate gibt es in der Regel spezielle Funktionen, die in den entsprechenden Paketen enthalten sind. Zum Beispiel kann die Funktion `readr::read_tsv()` für das Einlesen von TSV-Dateien (die mit einem Tabulator- oder TAB-Zeichen getrennt sind) verwendet werden. Die meisten Funktionen zum Einlesen von (Text)Daten haben eine Reihe von Optionen, mit denen Sie den Importprozess anpassen können. Zu den häufigsten Optionen gehören:

Funktionsargument	**Beschreibung**
header	Gibt an, ob die Datendatei eine Kopfzeile enthält
sep	Gibt das Trennzeichen zwischen den Werten an
na	Gibt an, wie fehlende Werte codiert werden sollen
dec	Gibt das Dezimaltrennzeichen an

7 Zusammenfassung

Es gibt verschiedene Möglichkeiten, Daten in R einzulesen. Die gebräuchlichste Methode ist die Verwendung von Funktionen. Für jedes gängige Datenformat gibt es eine entsprechende Funktion.