Einführung in die Datenanalyse mit R: Datenaufbereitung

Marco Wähner


Gebäude 37.03.03.14


marco.waehner@hhu.de


Sprechstunde nach Vereinbarung

Datenaufbereitung I (Kapitel 4)

Datenaufbereitung

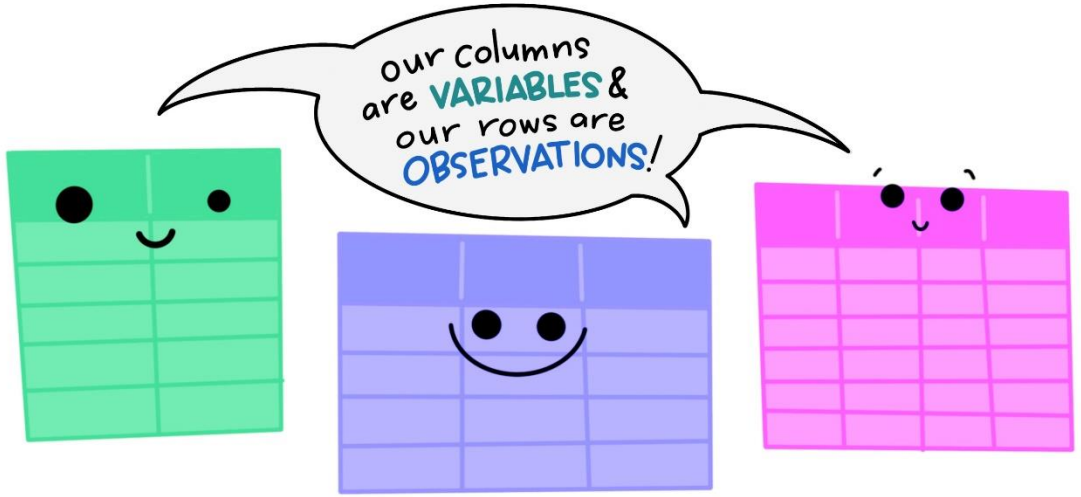
- 80/20 Ration – 80 Prozent Datenaufbereitung und 20 Prozent Datenauswertung

- Mögliche Schritte der Datenaufbereitung
  - Subsetting und Filtern (z.B. nur Befragte aus Westdeutschland)
  - Rekodierung von Variablen (z.B. Alter und Geburtskohorten)
  - Neue Variablen erstellen (z.B. SES)
  - Missing Values
  - Coercion
  - Neue Variablennamen vergeben (variablen_name_einer_bestimmten_variable = v1)

- Voraussetzung: Daten sind im „tidy“ Format
  - „wide“ vs. „long“ Format

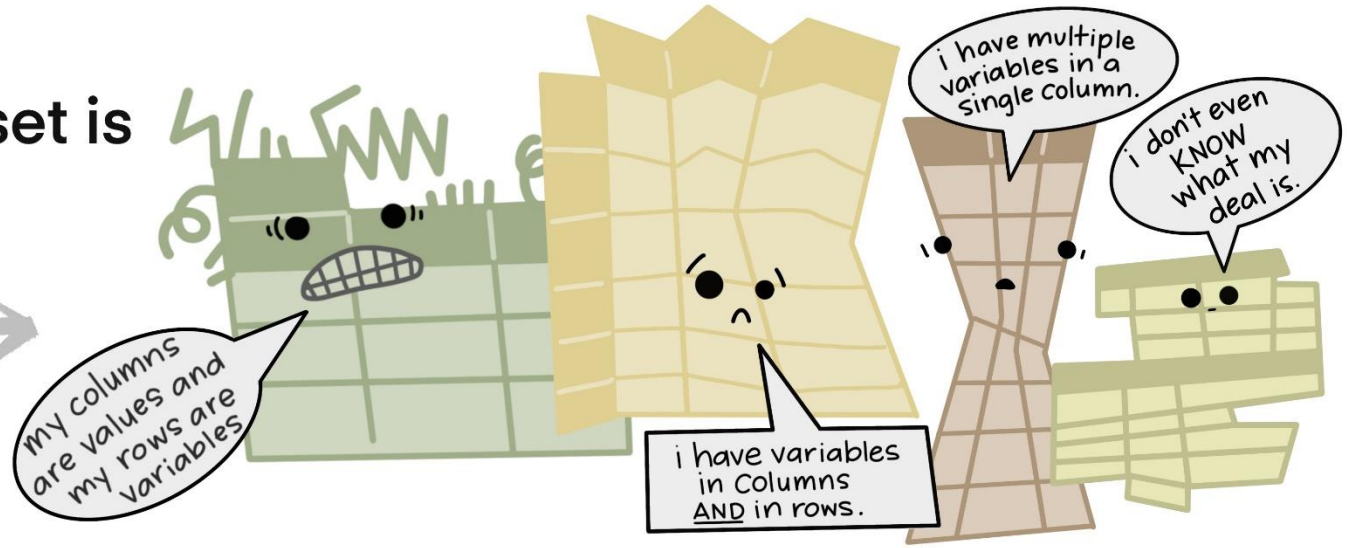
Tidy-Data

The standard structure of tidy data means that "tidy datasets are all alike..."



"...but every messy dataset is messy in its own way."

–HADLEY WICKHAM



https://github.com/allisonhorst/stats-illustrations

Tidy-Data

- Tidy-Data heißt
  - Jede Variable ist in einer Spalte
  - Jede Beobachtung ist in einer Zeile
  - Jeder Wert ist in einer Zelle

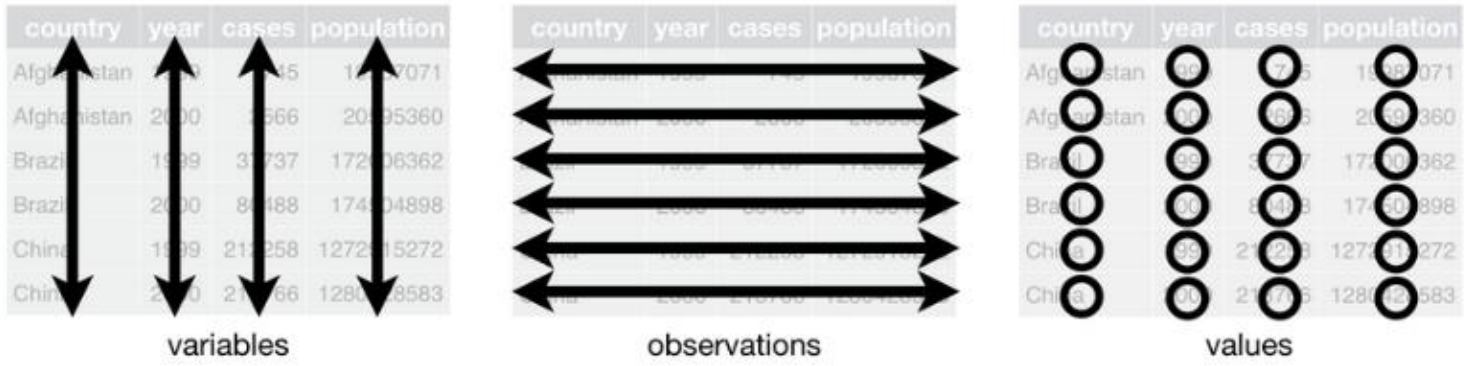
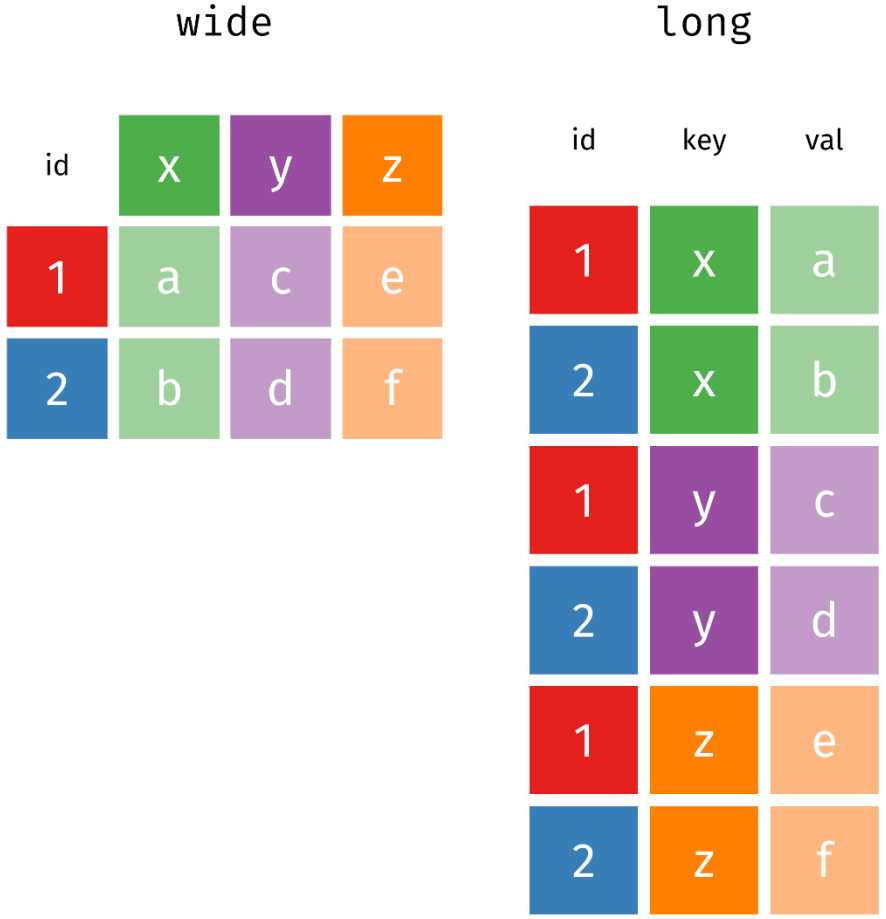


Figure 12.1: Following three rules makes a dataset tidy: variables are in columns, observations are in rows, and values are in cells. https://r4ds.had.co.nz/tidy-data.html

„wide“ vs. „long“ Format

- Interessant, wenn mit Panel oder Aggregat-Daten gearbeitet wird
- Unterschiedliche Funktionen zum aggregieren und disaggregieren von Daten



https://github.com/gadenbuie/tidyexplain#tidy-data

Datenaufbereitung

Base R

- Ohne zusätzliche Packages
- Funktioniert häufig über sog. Indexing [x] oder [[x]]

Tidyverse

- Muss zusätzlich installiert werden
- Funktioniert über den pipe-Operator %>%

- Es gibt Pro und Contra Argumente zum tidyverse

- Überblick über Variablen

```
#Überblick über Variablen verschaffen
attributes(allbus$eastwest)
```

```
## $label
## [1] "ERHEBUNGSGEBIET (WOHNGEBIET): WEST - OST"
##
## $format.spss
## [1] "F1.0"
##
## $display_width
## [1] 10
##
## $labels
## ALTE BUNDESLAENDER NEUE BUNDESLAENDER
##                  1                  2
```

- Überblick über Variablen

```
#Klasse der Variable
class(allbus$eastwest)
```

```
## [1] "numeric"
```

```
#fehlende Werte
sum(is.na(allbus$eastwest))
```

```
## [1] 0
```

Datenaufbereitung: Subsetting & Filtern

- Bedingungen formulieren über logische Operatoren

Operator	Beschreibung
<	Kleiner als
<=	Kleiner als oder gleich
>	Größer als
>=	Größer als oder gleich
==	Genau gleich
!=	Nicht genau gleich
| (x | y)	Oder (x oder y)
& (x & y)	Und (x und Y)

Praxis

Hands-On:

Data-Wrangling GLES 2021 (Lösung)

Kontakt

Vielen Dank!

Marco Wähner

 Gebäude 37.03.03.14

 marco.waehner@hhu.de

 Sprechstunde nach Vereinbarung