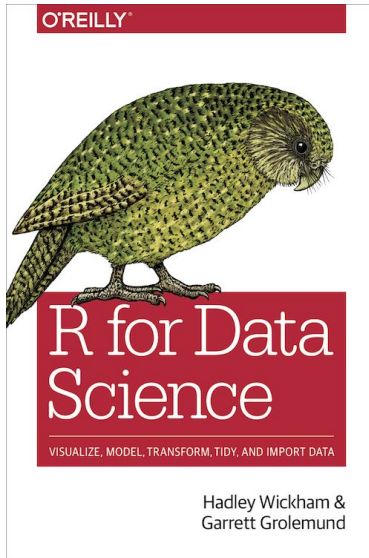




DATENTRANSFORMATION MIT R & dplyr

- Schritte in der Datenanalyse
- Erste Schritte mit R und RStudio
- Der Werkzeugkasten
- Daten laden mit `readr` und verwalten mit `tibble`
- Datentransformation mit dem `dplyr` Paket
 - Spalten auswählen
 - Zeilen filtern
 - Zeilen sortieren
 - Spalten verändern
 - Zeilen zusammenfassen
- Übungsaufgabe



Wickham, Hadley, and Garrett G. Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. First edition, O'Reilly, 2016. Online verfügbar: <https://r4ds.had.co.nz/>

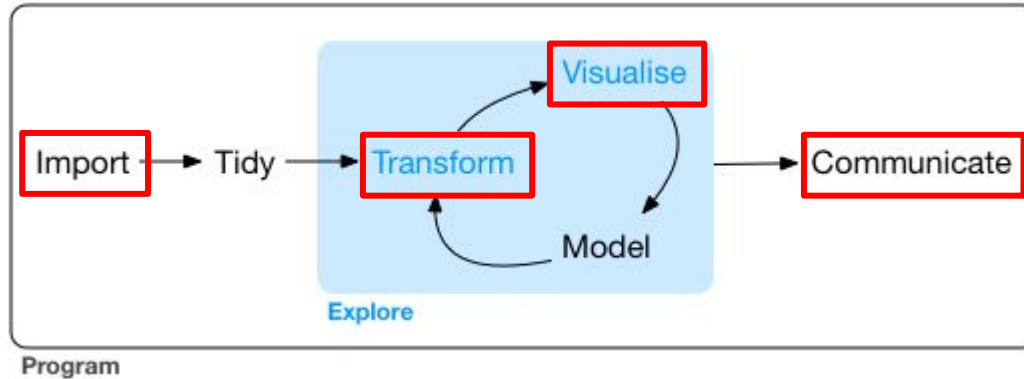
→ Kapitel 5 im Online-Buch / Kapitel 3 im deutschsprachigen PDF



Sauer, Sebastian. Moderne Datenanalyse mit R. Springer Gabler, 2019.

→ Kapitel 7

SCHRITTE IN DER DATENANALYSE



Source: Wickham, Hadley, and Garrett Grolemond. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. First edition, O'Reilly, 2016. URL: <https://r4ds.hadley.nz/exploratory-data-analysis.html>

Download, Installation R und RStudio

alternativ

Anmeldung und Login RStudio Cloud



Rundgang RStudio

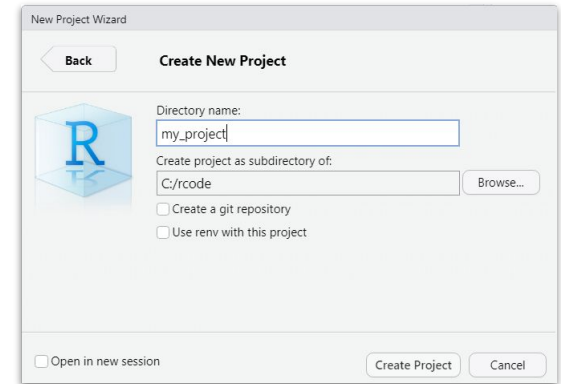
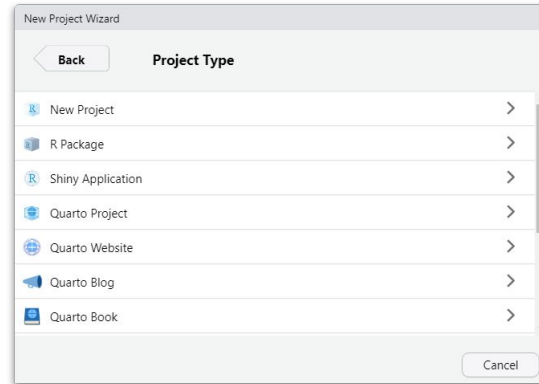
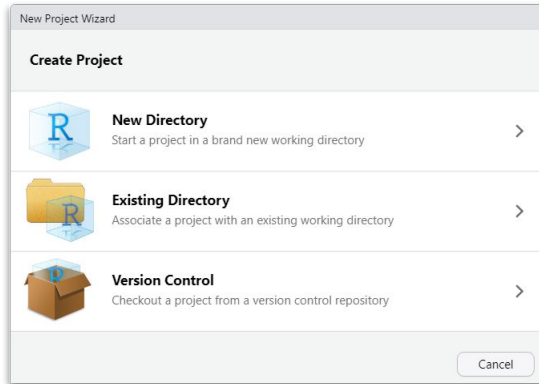
- Konsole und Skripteditor
- Pakete installieren
- Arbeitsverzeichnis
- Umgebung (Environment)
- Vorschaufenster
- Hilfe

ERSTE SCHRITTE MIT R UND RSTUDIO

PROJEKT ERSTELLEN

Projekte bündeln Einstellungen, Skripte und Arbeitsverzeichnis:

- Im RStudio → File → New Project
- Wählt “New Directory” als Option und dann “New Project”
- Wählt das übergeordnete Verzeichnis (z. B. **C:\rcode**) und gibt dem Projekt einen Namen
- Projekt wird dann in **C:\rcode\my_project** angelegt



- Daten laden u.a. mit `readr` oder `jsonlite`
- Daten verwalten mit `tibble`
- Daten transformieren mit `dplyr`
 - `select()`
 - `filter()`
 - `arrange()`
 - `mutate()`
 - `summarise()` / `group_by()`

- Arbeitsumgebung(en)
 - R
 - RStudio
 - Databricks (*für Big Data*)



arrange

Zeilen sortieren

select
Spalten auswählen

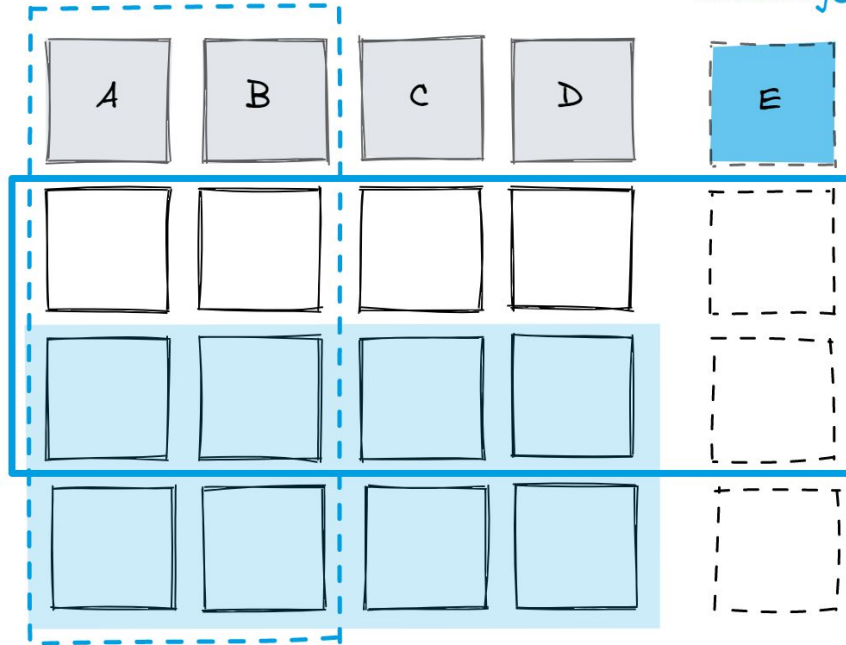
mutate
transmute
Neue Spalten
hinzufügen

group_by
summarise

Σ

Zeilen zusammenfassen
(aggregieren)

Zeilen filtern
filter



DATEN LADEN

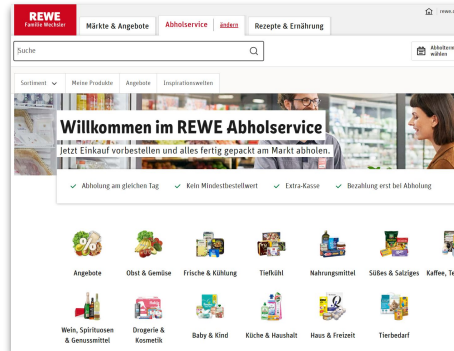
ÜBUNGSDATENSÄTZE

- Daten laden mit `{readr}` (Excel, CSV) oder `{jsonlite}` (JSON), oder `readRDS` (R-Format)
- `{janitor}` und `clean_names` für bessere Spaltennamen
- Datensätze als Einführungsbeispiele:

Campusbier Online-Verkäufe (CSV)



REWE-Online Produkte (CSV)



Tweets der Ampel-Regierung (JSON / RDS)



- Daten verwalten mit Dataframes und `{tibble}`
- Tibbles als moderne Dataframes
 - Verbesserte Ausgabe (z. B. `print(n = 10, width = Inf)`)
 - Keine automatische Konvertierung von Strings in Faktoren
 - Keine `rownames`
 - Variablennamen werden beim Erstellen nicht verändert

*Tibbles oder Dataframes? Beides sind
Tabellen wie aus Excel bekannt... nur in R*

- Bestimmte Spalten auswählen mit `{dplyr}`
 - `select()`
 - Nach Name
 - Nach Namensmuster (`starts_with`, `ends_with`, `contains`)
 - Spaltenbereich (`last_col`)
 - Alle oder einige aus einer Liste (`all_of`, `any_of`)
 - Mittels einer Funktion (`where()`)

- Zeilen einschränken mit `{dplyr}`
 - `filter()`
 - Einfache Bedingungen (`==`, `!=`, `<`, `>`)
 - Verknüpfte Bedingungen (`&`, `|`, `!`, `xor`)
 - Nach Mengenzugehörigkeit (`%in%`)
 - Fehlende Werte (`NA`, `is.na`)
 - Einfache Suche in Texten (`str_detect`)

- Ergebnis sortieren mit `{dplyr}`
 - `arrange()`
 - Aufsteigende Sortierung nach einer oder mehr Spalten
 - Absteigende Sortierung (`desc` oder `-`)

- Neue berechnete Spalten erzeugen mit `{dplyr}`
 - `mutate()`
 - Neue berechnete Spalten hinzufügen (+, -, /, *, %, ^, `paste0`)
 - Vektorisierte Aggregationen (`mean`, `sum`, `max`, `min`, `lag`, `lead`)
 - Nur beteiligte Spalten beibehalten (`.keep = "used"`)
 - Einfügestelle bestimmen mit `.before` und `.after`
 - `transmute()`
 - Neue Spalten hinzufügen und gleichzeitig alle anderen entfernen

- Daten zusammenfassen mit `{dplyr}`
 - `count()`, `tally()`, `distinct()` für schnelle Aggregation
 - `summarise()`
 - Daten aggregieren (`mean`, `median`, `quantile`, `sd`, `IQR`, `mad`, `sum`, `max`, `min`, `n`, `n_distinct`, `first`, `nth`, `last`)
 - `group_by()`
 - Bilden von Gruppen, nach denen aggregiert werden soll
 - Das `janitor` Paket für schnelle relative Anteile (%) und Kreuztabellen mit `tabyl()`

Ihr seid als Geschäftsführer neu im Campusbier-Projekt und sollt euch ein erstes Bild über das Geschäft machen. Alles, was ihr bekommt, sind die beiden Datensätze `orders.csv` und `line_items.csv`!

- Wie nähert ihr euch dem Datensatz?
- Überlegt euch zu zweit mindestens 3 Fragen, die ihr an die Daten stellen wollt! Schaut euch dazu die verfügbaren Spalten an!
- Erstellt R-Befehle, um die Fragen zu beantworten! (ohne Visualisierung)