

Statistik1

Daten umformen

Sebastian Sauer

Einstieg

Lernziele

- Sie können folgende Verben des Datenjudos anwenden: `arrange`, `filter`, `select`, `summarise`, `group_by`, `mutate`.
- Sie können R-Befehle mit der “Pfeife” verketteten.

Definition: Datenjudo

Definition: Datenjudo

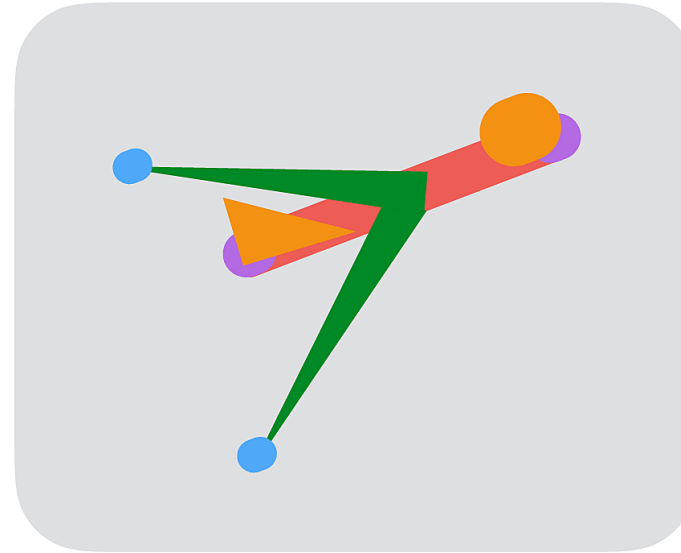
Mit *Datenjudo* meint man den Prozess des Aufbereitens, Umformens oder Zusammenfassens von Daten — sei es für einzelne Beobachtungen (Zeilen), Variablen (Spalten) oder eine ganze Datentabelle.

Praxisbezug: Data Science ist 80 % Aufbereiten

Laut *Harvard Business Review* verbringen Data Scientisten ca. 80 % ihrer Zeit mit dem *Aufbereiten* von Daten ([Bowne-Anderson, 2018](#)).

Uncoole Tätigkeiten wie Tippfehler entfernen oder Daten nutzbar machen — aber unverzichtbar. Das Aufbereiten von Daten verlangt keine aufwändige Mathematik, sondern ein paar Handgriffe und Kniffe: das *Datenjudo*.

Das Lego-Prinzip



Das Lego-Prinzip ([Sauer, 2019](#))

Komplexe Datenaufbereitung wird elegant, wenn man sie in einfache Teilschritte zergliedert — so wie ein komplexes Lego-Modell aus einfachen Bausteinen besteht.

Die Verben des Datenjudos

Ein einfacher Beispieldatensatz

Im R-Paket `dplyr` (Teil von `tidyverse`) gibt es eine Handvoll Verben, die das Datenjudo abdecken. Wir betrachten sie zunächst an einem kleinen Spielzeug-Datensatz:

id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

Die Verben wohnen im Paket `dplyr`, das mit `library(tidyverse)` gestartet wird.

Tabelle sortieren: **arrange**

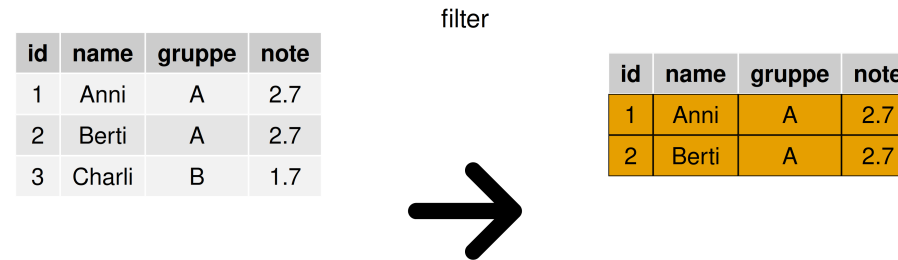
arrange			
id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7



id	name	gruppe	note
3	Charli	B	1.7
1	Anni	A	2.7
2	Berti	A	2.7

```
1 arrange(mariokart, total_pr) # aufsteigend (Voreinstellung)
2 arrange(mariokart, -total_pr) # absteigend
```

Zeilen filtern: `filter`



```
1 filter(mariokart, stock_photo == "yes" & cond == "new") # UND
2 filter(mariokart, stock_photo == "yes" | cond == "new") # ODER
```

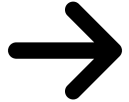
Spalten auswählen: **select**

id	name	gruppe	note	select	id	note
1	Anni	A	2.7		1	2.7
2	Berti	A	2.7		2	2.7
3	Charli	B	1.7		3	1.7

```
1 select(mariokart, cond, total_pr)
2 select(mariokart, 2:5)      # Spalten 2 bis 5
3 select(mariokart, -1)      # alle Spalten außer Spalte 1
```

Spalten zusammenfassen: `summarise`

summarise			
id	name	gruppe	note
1	Anni	A	2.7
2	Berti	A	2.7
3	Charli	B	1.7

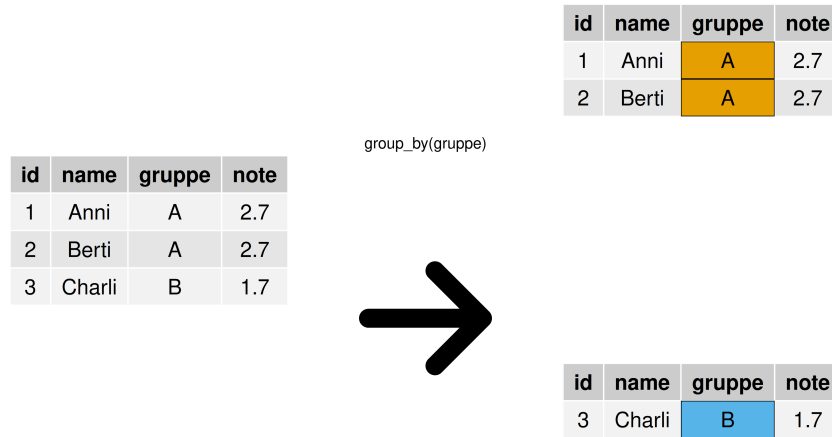


note_mw
2.4

```
1 mariokart_mittelwert <- summarise(mariokart,  
2                                   preis_mw = mean(total_pr))
```

`summarise` fasst eine *Spalte* zu einer einzelnen *Zahl* zusammen.

Tabelle gruppieren: `group_by`



```
1 mariokart_gruppiert <- group_by(mariokart, cond)  
2 summarise(mariokart_gruppiert, preis_mw = mean(total_pr))
```

Gruppieren teilt die Tabelle in Teiltabellen — nachfolgende Berechnungen laufen *für jede Teiltabelle einzeln*.

Spalten verändern: **mutate**

				mutate					
id	name	gruppe	note		id	name	gruppe	note	punkte
1	Anni	A	2.7	➔	1	Anni	A	2.7	73
2	Berti	A	2.7		2	Berti	A	2.7	72
3	Charli	B	1.7		3	Charli	B	1.7	89

```
1 mariokart_yen <- mutate(mariokart, total_pr_yen = total_pr * 133)
```

mutate berechnet eine (neue oder bestehende) Spalte für jede Zeile neu — z. B. addieren, multiplizieren, transformieren.

Zeilen zählen: **count**

Bei nominalskalierten Daten reicht oft schon das Zählen der Zeilen:

```
1 mariocart_counted <- count(mariokart, cond)
2 mutate(mariocart_counted, Anteil = n / sum(n))
```

Liefert Häufigkeiten je Ausprägung — und mit **mutate** gleich noch den Anteil.

Verben am Fließband

Tidyverse-Befehle erwarten typischerweise eine Tabelle als Input und liefern eine Tabelle als Output zurück.



Die Pfeife

Definition: Pfeife

Definition: Pfeife

“Und dann” heißt auf Errisch `%>%` oder (synonym) `|>`. Man nennt diesen Befehl “Pfeife” (engl. *pipe*).

Die Pfeife `%>%` wohnt im Paket `tidyverse` (genauer: `magrittr`). Seit R 4.1 gibt es auch eine eingebaute Pfeife: `|>`.

Russische Puppen: verschachtelte Befehle

R arbeitet einen verschachtelten Befehl *von innen nach außen* ab:

1. `sum(x - mean(x))`
2. `sum(x - 2)`
3. `sum(-1, 0, 1)`
4. `0`

Bei mehr Verschachtelung wird es schnell unübersichtlich:

```
1 fasse_zusammen(  
2   gruppieren(  
3     wähle_spalten(  
4       filter_zeilen(meine_daten))))
```

Die Pfeife zur Rettung

Statt verschachtelt schreibt man mit der Pfeife die Schritte nacheinander:



```

1 mariokart %>%
2   filter(total_pr < 100) %>%
3   select(cond, total_pr) %>%
4   group_by(cond) %>%
5   summarise(total_pr_mean = mean(total_pr))
  
```

`filter(mariokart, x)` entspricht `mariokart |> filter(x)`. Allgemein: `d |> f(x) = f(d, x)`.

Beispiele für Forschungsfragen

Das teuerste Spiel?

```
1 mariokart %>%  
2   filter(stock_photo == "yes") %>%  
3   group_by(cond) %>%  
4   summarise(total_pr_max = max(total_pr))
```

Das teuerste Spiel mit Foto kostet 75 Dollar (neu) bzw. 62 Dollar (gebraucht).

Die mittlere Versandpauschale?

Getrennt nach Anzahl der Lenkräder, nur Gruppen mit mind. 10 Spielen:

```
1 mariokart %>%  
2   filter(wheels < 3) %>%  
3   group_by(wheels) %>%  
4   summarise(mittlere_versandkosten = mean(ship_pr),  
5             anzahl_spiele = n())
```

Die mittleren Versandkosten liegen zwischen 2.7 und 3.6 Dollar, je nach Anzahl der Lenkräder.

Verkaufspreis in Yen?

Mittlerer Verkaufspreis in Yen, minus 10 % Provision:

```
1 mariokart %>%  
2   select(total_pr) %>%  
3   mutate(total_pr_yen = total_pr * 133) %>%  
4   summarise(  
5     preis_yen_mw = mean(total_pr_yen),  
6     preis_yen_mw_minus_10proz = preis_yen_mw - 0.1 * preis_yen_mw)
```

`summarise` kann auch mehrere Berechnungen gleichzeitig ausführen.

Praxisbezug

Datenjudo in der Forschung: COVIDiSTRESS

Mulukom et al. ([2020](#)) untersuchten psychologische Auswirkungen der Covid19-Pandemie (Vereinsamung, Angst, Depression) — angemeldet unter der Projekt-ID *tsjnb* bei der Open Science Foundation.

Die R-Syntax zur Datenaufbereitung ist öffentlich einsehbar — ein guter Teil davon entspricht dem, was Sie in diesem Kapitel gelernt haben. Ein Beispiel dafür, wie angewandte Forschung und ein akutes gesellschaftliches Problem zusammenkommen.

Wie man mit Statistik lügt

Der ungewichtete-Mittelwert-Trick

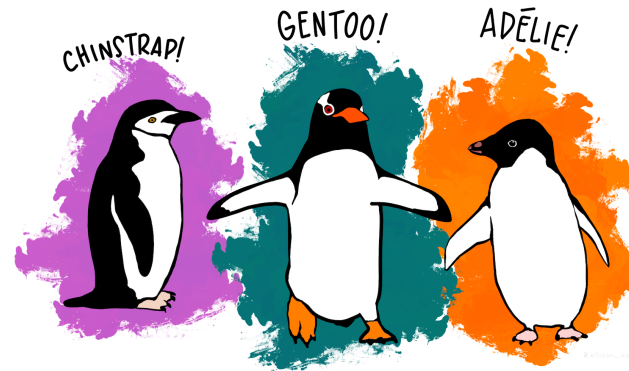
Eine Analystin möchte zeigen, dass der Verkaufspreis “viel zu niedrig” ist (Ist-Wert: 50 Euro):

```
1 mariokart_wheels <- mariokart %>%  
2   group_by(wheels) %>%  
3   summarise(pr_mean = mean(total_pr), count_n = n())  
4  
5 mariokart_wheels %>%  
6   summarise(mean(pr_mean))    # ungewichteter Mittelwert -> 56 Euro
```

Der *ungewichtete* Mittelwert über die Gruppen (56 Euro) ist falsch und irreführend — der korrekte, *gewichtete* Mittelwert ergibt weiterhin 50 Euro.

Fallstudien

Die Pinguine



Possierlich: Die Pinguine ([Horst, 2024](#))

Forschungsfragen zum Datensatz `penguins` (Zielvariable: Gewicht):

1. Medianes Gewicht, gruppiert nach Spezies?
2. Wie viele Pinguine pro Spezies?
3. Schwerster und leichtester Pinguin pro Spezies?

Fallstudie COVIDiSTRESS



Studie COVIDiSTRESS ([Lieberoth et al., 2022](#))

Internationale Erhebung zu psychologischem Stress, Verhaltensregeln und Vertrauen in Institutionen während der Covid19-Pandemie — Daten frei verfügbar zur eigenständigen Analyse.

Zusammenfassung

- **Datenjudo:** Aufbereiten, Umformen, Zusammenfassen von Daten
- Sechs Kern-Verben: `arrange`, `filter`, `select`, `summarise`, `group_by`, `mutate` (+ `count`)
- Jedes Verb: Tabelle rein, Tabelle raus
- **Pfeife** (`%>%` bzw. `|>`) verkettet Befehle statt sie zu verschachteln: `d |> f(x) = f(d, x)`
- Vorsicht bei Mittelwerten über Gruppen: ungewichtet $\neq$ gewichtet

Literaturhinweise

- Sauer ([2019](#)), Kap. 7 — Einführung in die Datenaufbereitung mit R
- Wickham & Grolemund ([2018](#)), Kap. 5 — vertieft die Themen dieses Kapitels

Referenzen

- Bowne-Anderson, H. (2018). What Data Scientists Really Do, According to 35 Data Scientists. *Harvard Business Review*. <https://hbr.org/2018/08/what-data-scientists-really-do-according-to-35-data-scientists>
- Horst, A. (2024). *Statistics Artwork* [Artwork]. <https://allisonhorst.com/>
- Lieberoth, A., Rasmussen, J., Stoeckli, S., Tran, T., Čepulić, D.-B., Han, H., Lin, S.-Y., Tuominen, J., Travaglino, G., & Vestergren, S. (2022). *COVIDiSTRESS Global Survey*. <https://doi.org/10.17605/OSF.IO/Z39US>
- Mulukom, V. van, Muzzulini, B., Rutjens, B., Lissa, C. J. van, & Farias, M. (2020). *Psychological Impact of COVID-19 Pandemic*. <https://doi.org/10.17605/OSF.IO/TSJNB>
- Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren*. Springer. <https://www.springer.com/de/book/9783658215866>
- Wickham, H., & Grolemund, G. (2018). *R für Data Science: Daten importieren, bereinigen, umformen, modellieren und visualisieren* (F. Langenau, Übers.). O'Reilly. <https://r4ds.had.co.nz/index.html>