

# rstatsZH - Data Science mit R

## Daten Transformation mit dplyr - Teil 2

Lars Schöbitz

2022-09-22

# Rückblick - Woche 3

## ggplot2 Visualisierungen anpassen

- Skalierungen
  - `scale_` Funktionen
- Aussehen
  - `theme` Funktionen

## dplyr Funktionen

- `filter`: Zeilen mit bestimmten Kriterien auswählen

# dplyr::filter() - Zeilen mit bestimmten Kriterien auswählen

**dplyr:: filter()**

KEEP ROWS THAT  
s.a.t.i.s.f.y  
*your CONDITIONS*

keep rows from... this data... ONLY IF... type is "otter" AND site is "bay"

```
filter(df, type == "otter" & site == "bay")
```

The illustration features a smiling orange circular character with a simple face and a small tuft of hair. It is pointing its right hand towards a map of a coastal area with a blue ocean and green landmasses. A small brown otter is swimming in the water near a rocky shore labeled 'BAY'. The map includes a compass rose indicating North. To the right of the map is a table titled 'filter(df, type == "otter" & site == "bay")'. The table has columns 'type', 'food', and 'site'. The rows are:

type	food	site
otter	urchin	bay
Shark	seal	channel
otter	abalone	bay
otter	crab	wharf

Below the table is the handle '@allison\_horst'.

To the right of the table are two cartoon otters. The purple otter is standing on top of a green spherical object, looking happy with a checkmark (✓) next to it. The green otter is sitting on the ground, looking sad with a red X next to it. This visualizes how the filter function keeps rows where both conditions are true (the purple otter's row) and removes rows where either condition fails (the green otter's row).

# Ziele für diese Woche

Am Ende dieser Woche könnt ihr:

- mehr als zehn Funktionen des R Package `{dplyr}` anwenden, um
  - Daten einzugrenzen
  - neue Variablen zu erstellen
  - zusammenfassende Statistiken zu berechnen
- `NA` Werte aus euren Daten entfernen

# Hausaufgabe 3 - Lösungen

- **GitHub Organisation:** rstatsZH
  - <https://github.com/rstatsZH/>
- **Repo:** ha-03-hallo-dplyr
  - <https://github.com/rstatsZH/ha-03-hallo-dplyr>
- **R Markdown Datei:** ha-03-solutions.Rmd
  - <https://github.com/rstatsZH/ha-03-hallo-dplyr/blob/main/ha-03-solutions.Rmd>

# Praktikum 5 - dplyr

## Live Code

1. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
2. **RStudio Cloud / Projects:** Öffne erneut das Praktikum 05
3. Folgt wieder auf dem Bildschirm

# Praktikum 6 - dplyr

In 2er Teams

1. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
2. **RStudio Cloud / Projects:** Öffne erneut das Praktikum 06

# dplyr::select() - Spalten beim Namen auswählen

```
# Erste Möglichkeit  
penguins %>%  
  select(species, bill_length_mm, bill_depth_mm, flipper_length_mm)  
  
# Zweite Möglichkeit  
penguins %>%  
  select(species, bill_length_mm:flipper_length_mm)  
  
# Dritte Möglichkeit - nicht zu empfehlen  
penguins %>%  
  select(1, 3, 4, 5)
```

```
# A tibble: 4 × 4  
species bill_length_mm bill_depth_mm flipper_length_mm  
  <fct>      <dbl>        <dbl>          <int>  
1 Adelie     39.1         18.7           181  
2 Adelie     39.5         17.4           186  
3 Adelie     40.3         18              195  
4 Adelie      NA            NA             NA
```

## dplyr::arrange() - Zeilen neu ordnen

```
# Aufsteigende Reihenfolge  
penguins %>%  
  arrange(body_mass_g)  
  
# Absteigende Reihenfolge  
penguins %>%  
  arrange(desc(body_mass_g))
```

# dplyr::select() - Hilfsfunktionen

```
penguins %>%  
  select(starts_with(match = "BILL",  
                     # Ignoriere Gross- und Kleinschreibung  
                     ignore.case = TRUE)) # Standardeinstellung
```

```
# A tibble: 344 × 2  
  bill_length_mm bill_depth_mm  
        <dbl>         <dbl>  
1       39.1          18.7  
2       39.5          17.4  
3       40.3           18  
4        NA            NA  
5       36.7          19.3  
6       39.3          20.6  
# ... with 338 more rows
```

# dplyr::select() - Hilfsfunktionen

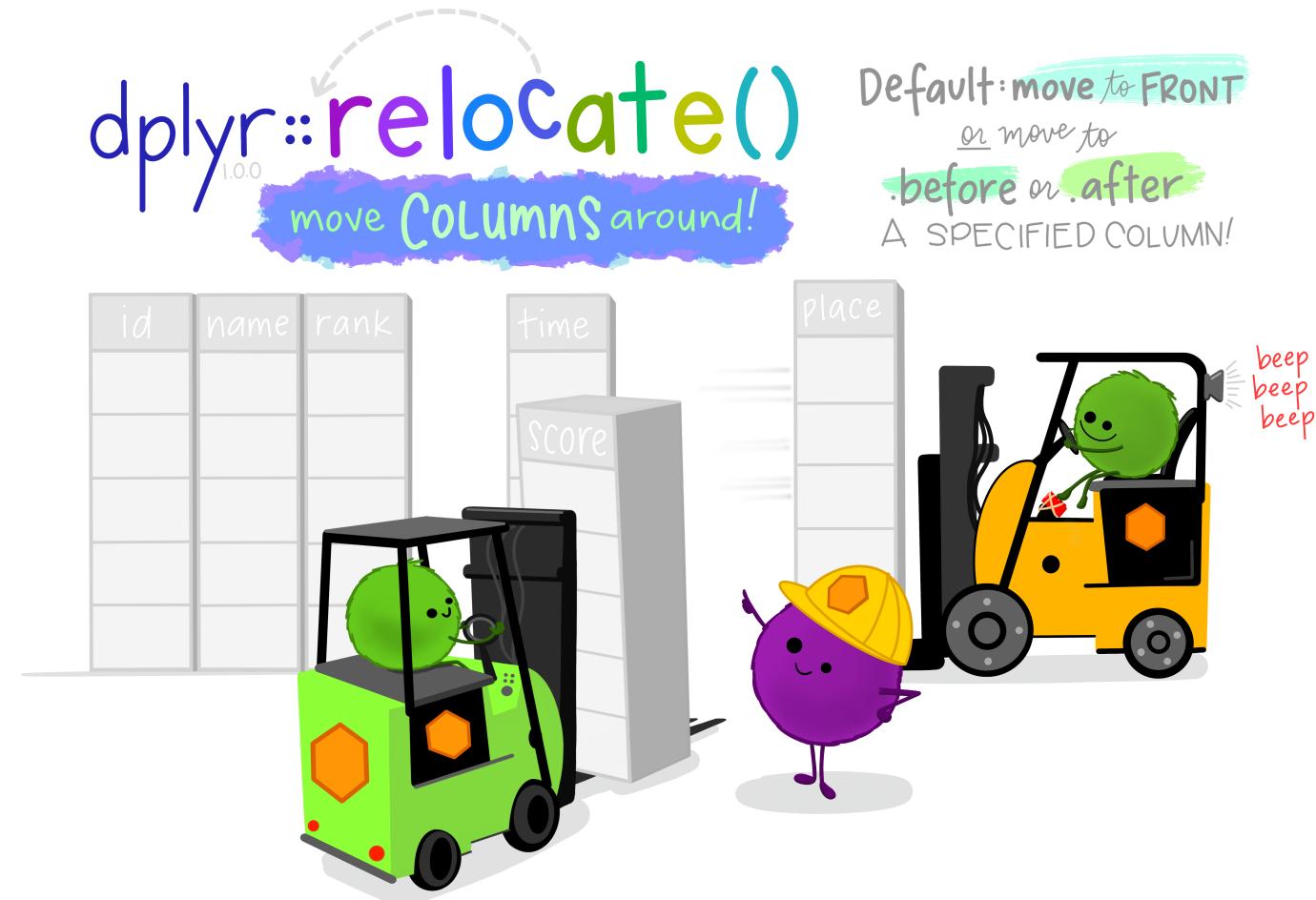
```
penguins %>%  
  select(starts_with(match = "BILL",  
                     # Ignoriere Gross- und Kleinschreibung  
                     ignore.case = FALSE)) # Geändert auf FALSE
```

```
# A tibble: 344 × 0
```

Mehr Informationen zu Hilfsfunktionen:

<https://dplyr.tidyverse.org/reference/select.html>

# dplyr::relocate() - Spalten verschieben



@allison\_horst

# dplyr::rename() - Spalten umbenennen

```
penguins %>%  
  rename(mass = body_mass_g) # neuer name = alter name
```

```
# A tibble: 344 × 8  
  species island    bill_leng...¹ bill_...² flipp...³   mass sex   year  
  <fct>   <fct>        <dbl>    <dbl>    <int> <int> <fct> <int>  
1 Adelie  Torgersen     39.1     18.7     181  3750 male  2007  
2 Adelie  Torgersen     39.5     17.4     186  3800 fema... 2007  
3 Adelie  Torgersen     40.3      18       195  3250 fema... 2007  
4 Adelie  Torgersen      NA       NA       NA    NA <NA>  2007  
5 Adelie  Torgersen     36.7     19.3     193  3450 fema... 2007  
6 Adelie  Torgersen     39.3     20.6     190  3650 male  2007  
# ... with 338 more rows, and abbreviated variable names  
#   `¹bill_length_mm`, `²bill_depth_mm`, `³flipper_length_mm`
```

# Praktikum 5 - dplyr

## Live Code

1. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
2. **RStudio Cloud / Projects:** Öffne erneut das Praktikum 05
3. Folgt wieder auf dem Bildschirm

# Praktikum 6 - dplyr - Teil 5

## Hausaufgabe

1. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
2. **RStudio Cloud / Projects:** Öffne erneut das Praktikum 06

# dplyr::mutate() - Spalten hinzufügen



# dplyr::mutate() - Mit anderen Funktionen

Factor relevel    Plot Code    Plot

```
library(forcats)

penguins_relevel <- penguins %>%
  mutate(species = fct_relevel(species, c("Chinstrap", "Gentoo", "Adelie")))
```

# Praktikum 6 - Lösungen

- **GitHub Organisation:** rstatsZH
  - <https://github.com/rstatsZH/>
- **Repo:** prak-06-wrangle-dplyr
  - <https://github.com/rstatsZH/prak-06-wrangle-dplyr>
- **R Markdown Datei:** prak-06-wrangle-dplyr.Rmd
  - <https://github.com/rstatsZH/prak-06-wrangle-dplyr/blob/main/prak-06-solutions.Rmd>

# Praktikum 7 - dplyr

## Live Code

1. **E-Mail:** Öffne deine Email und klicke auf den Link zu deinem persönlichen GitHub repo für **prak-07**
2. **GitHub:** Klicke auf den grünen Button "Code" und kopiere den Link für das Repo in deine Zwischenablage
3. **RStudio Cloud:** Öffne deinen Arbeitsbereich für den Kurs in der RStudio Cloud
4. **RStudio Cloud / Projects:** Klicke auf "New Project from GitHub Repository"
5. **RStudio:** Finde den Datei Manager und Git
6. **Zoom Chat:** Schreibt wenn ihr soweit seid

# Feedback

# Ziele erreicht?

Bitte ausfüllen: [kutt.it/rstatszh-eval](https://kutt.it/rstatszh-eval)



# Hausaufgabe



Danke

Für die Aufmerksamkeit!

Für die R packages `{xaringan}` und `{xaringanthemer}` mit welchen die Folien geschrieben wurden.

Eine PDF Version der Folien kann hier heruntergeladen werden:

[https://github.com/rstatsZH/website/raw/master/slides/e1\\_d04-data-transform-teil2/e1\\_d04-data-transform-teil2.pdf](https://github.com/rstatsZH/website/raw/master/slides/e1_d04-data-transform-teil2/e1_d04-data-transform-teil2.pdf)

---

Für Data Science in a Box und Remaster the Tidyverse, von welchen ich Materialien für diesen Kurs nutze und welche genau wie diese Folien mit Creative Commons Attribution Share Alike 4.0 International lizenziert sind.