

# rstatsZH - Data Science mit R

## Daten Transformation mit dplyr - Teil 2

Lars Schöbitz

2020-03-29

# Rückblick - Woche 3

## ggplot2 Visualisierungen anpassen

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

## dplyr Funktionen

- `filter`: Zeilen mit bestimmten Kriterien auswählen
- `arrange`: Zeilen neu ordnen
- `select`: Spalten beim Namen 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. The map shows a blue ocean, a green landmass, and a body of water labeled 'BAY'. A north arrow is visible. To the right of the map is a data table with three rows highlighted in purple. Below the table is a small caption: '@allison\_horst'. To the right of the table are two cartoon animals: a purple otter-like creature standing on top of a green spherical character.

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

# 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 x 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))
```

# 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 - Hallo dplyr

1. **RStudio Cloud:** Öffnet den Arbeitsbereich für den Kurs
2. **RStudio Cloud - Projects:** Öffnet das Projekt für Hausaufgabe 3
3. **File-Manager:** Öffnet eure R Markdown Datei (ha-03.Rmd) für die Hausaufgabe 3
4. Strickt das Dokument

**Wo gibt es Klärungsbedarf?**

# 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() - Hilfsfunktionen

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

```
# A tibble: 344 x 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

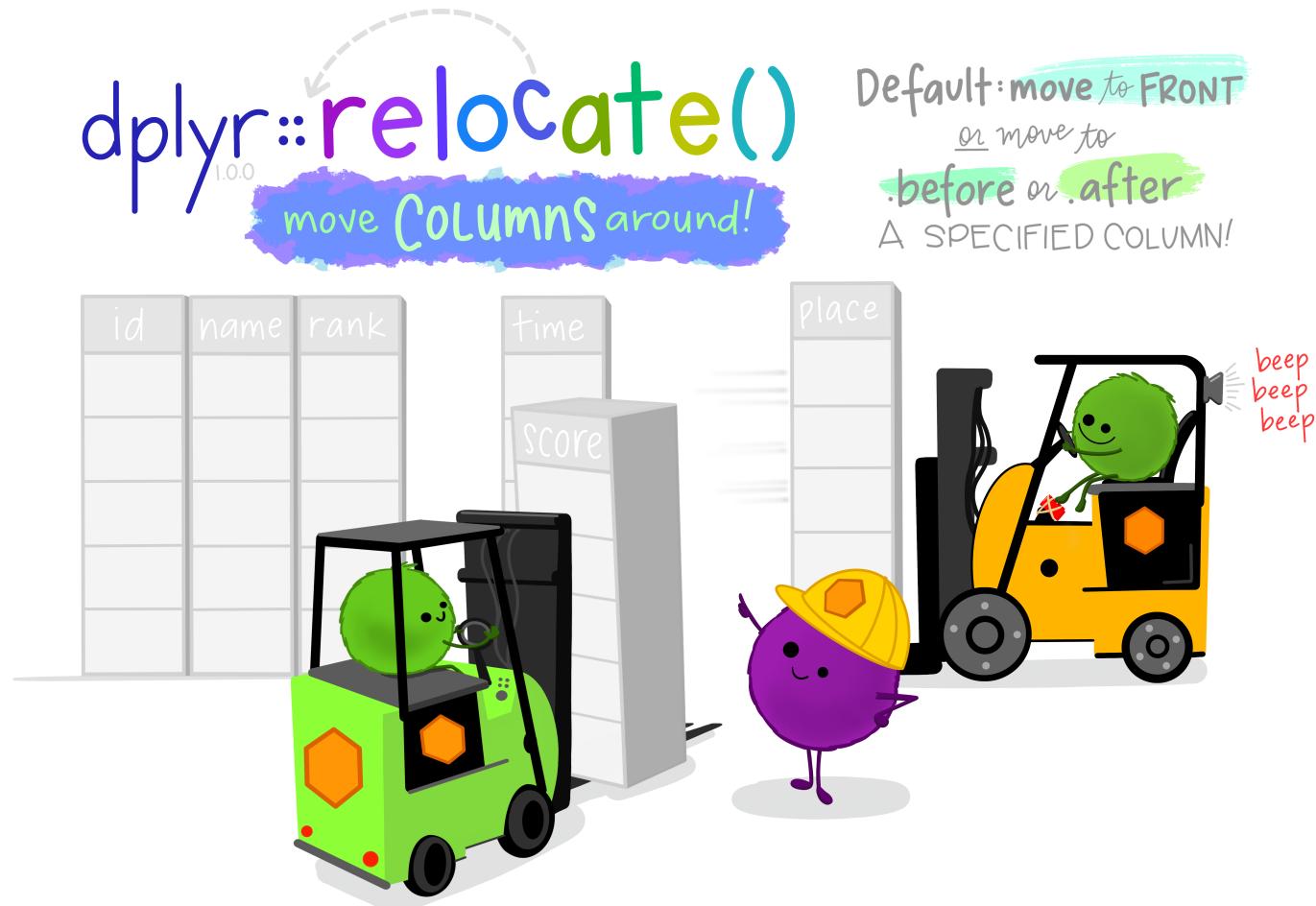
```
penguins %>%  
  select(starts_with(match = "BILL",  
                     # Ignoriere Gross- und Kleinschreibung  
                     ignore.case = FALSE)) # Geändert auf FALSE
```

```
# A tibble: 344 x 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 x 8  
  species island    bill_length_mm bill_depth_mm flipper_length_...  
  <fct>   <fct>        <dbl>        <dbl>            <int>  
1 Adelie  Torgersen     39.1       18.7             181  
2 Adelie  Torgersen     39.5       17.4             186  
3 Adelie  Torgersen     40.3        18              195  
4 Adelie  Torgersen      NA          NA              NA  
5 Adelie  Torgersen     36.7       19.3             193  
6 Adelie  Torgersen     39.3       20.6             190  
# ... with 338 more rows, and 3 more variables: mass <int>,  
#       sex <fct>, year <int>
```

# 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::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 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.