rstatsZH - Data Science mit R

Mit mehreren Dataframes arbeiten / Tabellen darstellen

Lars Schöbitz 2021-07-08

Rückblick - Woche 6

- Erweiterte Vektoren
 - Faktoren
 - Datums- und Zeitwerte
 - Tibbles
- Daten importieren
 - o read_csv()
 - o read_excel()
- Tidy Data Konzept
- Daten drehen (pivoting) mit {tidyr}
 - o pivot_longer()
 - o pivot_wider()

Hausaufgabe 6

Hausaufgabe 6



Hausaufgabe 6 - Lösungen

- **GitHub Organisation:** rstatsZH
 - https://github.com/rstatsZH/
- Repo: ha-06-treibhausgase
 - https://github.com/rstatsZH/ha-06-treibhausgase
- R Markdown Datei: ha-06-solutions.Rmd
 - https://github.com/rstatsZH/ha-06-treibhausgase/blob/main/ha-06-solutions.Rmd

Ziele für diese Woche

Am Ende dieser Woche könnt ihr:

- Mehrere Dataframes zusammenfügen
- Tabellen mit verschiedenen Packages darstellen erstellen
- Eine eigene Funktion für ein {ggplot2} theme schreiben
- Mit den im Kurs erlernten Fähigkeiten selbstständig weiter arbeiten

Mit mehreren Dataframes arbeiten

Wir...

haben mehrere Dataframes

WO len diese zusammenbringen

Data: Women in science

Informationen zu 10 Frauen in der Wissenschaft welche die Welt verändert haben

name

Ada Lovelace

Marie Curie

Janaki Ammal

Chien-Shiung Wu

Katherine Johnson

Rosalind Franklin

Vera Rubin

Gladys West

Flossie Wong-Staal

Jennifer Doudna

Quelle: Discover Magazine

Inputs - Drei Dataframes

professions dates works

```
# A tibble: 10 x 2
                    profession
  name
  <chr>
                    <chr>
1 Ada Lovelace
                   Mathematician
                   Physicist and Chemist
2 Marie Curie
3 Janaki Ammal Botanist
4 Chien-Shiung Wu Physicist
5 Katherine Johnson Mathematician
6 Rosalind Franklin Chemist
7 Vera Rubin Astronomer
8 Gladys West Mathematician
9 Flossie Wong-Staal Virologist and Molecular Biologist
10 Jennifer Doudna Biochemist
```

Gewünschter Output

```
# A tibble: 10 x 5
  name profession
                          birth year death year known for
  <chr> <chr>
                               <dbl> <dbl> <chr>
1 Ada Lov... Mathematician
                                             NA first computer a...
                                  NA
2 Marie C... Physicist an...
                                             NA theory of radioa...
                                  NA
3 Janaki ... Botanist
                                           1984 hybrid species, ...
                                1897
4 Chien-S... Physicist
                                           1997 confim and refin...
                            1912
 5 Katheri... Mathematician
                                1918
                                           2020 calculations of ...
6 Rosalin... Chemist
                                1920
                                           1958 <NA>
                                1928
                                           2016 existence of dar...
7 Vera Ru... Astronomer
8 Gladys ... Mathematician
                                1930
                                             NA mathematical mod...
 9 Flossie... Virologist a...
                                1947
                                             NA first scientist ...
10 Jennife... Biochemist
                                1964
                                             NA one of the prima...
```

Inputs als Erinnerung

names(professions)		nrow(professions)
[1] "name"	"profession"	[1] 10
names(dates)		nrow(dates)
[1] "name"	<pre>"birth_year" "death_y</pre>	[1] 8
names(works)		nrow(works)
[1] "name"	"known_for"	[1] 9

Dataframes zusammenfügen

Dataframes zusammenfügen

```
abcd_join(x, y)
```

- left_join(): alle Reihen aus x
- right_join(): alle Reihen aus y
- full_join(): alle Reihen aus x und y
- ...

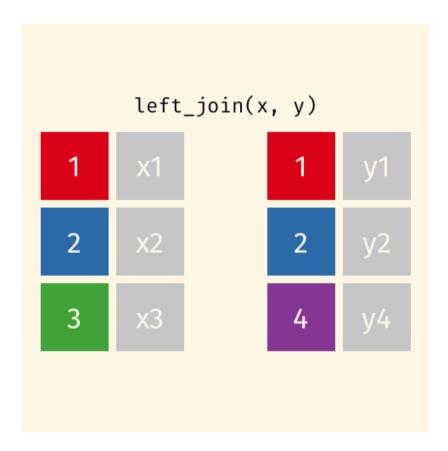
Beispiel

Für die nächsten Folien

```
# A tibble: 3 x 2
    id var_x
    <dbl> <chr>
1     1 x1
2     2 x2
3     3 x3
```

```
# A tibble: 3 x 2
    id var_y
    <dbl> <chr>
1     1 y1
2     2 y2
3     4 y4
```

left_join()



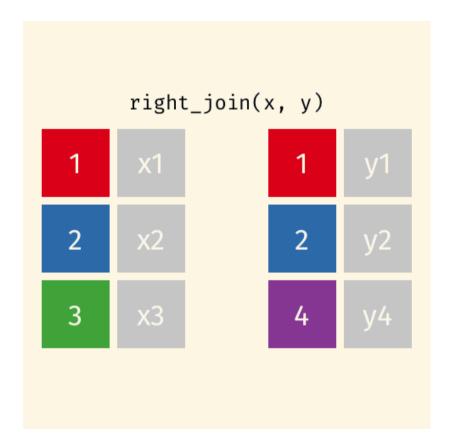
```
left_join(tib_x, tib_y)
```

left_join()

```
professions %>%
   left_join(dates)
```

```
# A tibble: 10 x 4
                profession
                                        birth_year death_year
  name
  <chr> <chr>
                                             <fdb> <fdb>
1 Ada Lovelace Mathematician
                                                NA
                                                          NΑ
2 Marie Curie Physicist and Chemist
                                                NΑ
                                                          NA
3 Janaki Ammal Botanist
                                              1897
                                                        1984
4 Chien-Shiung ... Physicist
                                              1912
                                                        1997
5 Katherine Joh... Mathematician
                                              1918
                                                        2020
6 Rosalind Fran... Chemist
                                              1920
                                                        1958
7 Vera Rubin Astronomer
                                              1928
                                                        2016
8 Gladys West Mathematician
                                              1930
                                                          NA
9 Flossie Wong-... Virologist and Molecular...
                                              1947
                                                          NA
10 Jennifer Doud... Biochemist
                                              1964
                                                          NA
```

right_join()



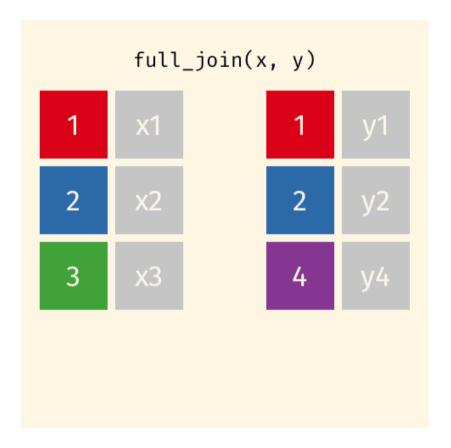
```
right_join(tib_x, tib_y)
```

right_join()

```
professions %>%
    right_join(dates)
```

```
# A tibble: 8 x 4
 name profession
                                       birth_year death_year
 <chr> <chr>
                                            <dbl>
1 Janaki Ammal Botanist
                                             1897
                                                      1984
2 Chien-Shiung ... Physicist
                                             1912
                                                      1997
3 Katherine Joh... Mathematician
                                             1918
                                                      2020
4 Rosalind Fran... Chemist
                                             1920
                                                      1958
5 Vera Rubin Astronomer
                                             1928
                                                       2016
6 Gladys West Mathematician
                                             1930
                                                        NA
7 Flossie Wong-... Virologist and Molecular ...
                                             1947
                                                        NA
8 Jennifer Doud... Biochemist
                                             1964
                                                        NA
```

full_join()



```
full_join(tib_x, tib_y)
```

full_join()

```
dates %>%
  full_join(works)
```

```
# A tibble: 10 x 4
             birth_year death_year known_for
  name
                  <dbl> <dbl> <chr>
  <chr>
1 Janaki Am...
                  1897
                             1984 hybrid species, biodiversity...
2 Chien-Shi... 1912
                             1997 confim and refine theory of ...
                             2020 calculations of orbital mech...
3 Katherine... 1918
4 Rosalind ... 1920
                             1958 < NA>
5 Vera Rubin 1928
                             2016 existence of dark matter
                               NA mathematical modeling of the...
6 Gladys We...
                  1930
7 Flossie W...
                  1947
                               NA first scientist to clone HIV...
8 Jennifer ... 1964
                               NA one of the primary developer...
9 Ada Lovel...
                               NA first computer algorithm
                  NA
                               NA theory of radioactivity, di...
10 Marie Cur...
                    NA
```

Alles in einer Code Sequenz

```
professions %>%
  left_join(dates) %>%
  left_join(works)
```

```
# A tibble: 10 x 5
  name profession birth year death year known for
  <chr> <chr>
                               <dbl> <dbl> <chr>
1 Ada Lov... Mathematician
                                 NA
                                             NA first computer a...
2 Marie C... Physicist an...
                                 NA
                                             NA theory of radioa...
3 Janaki ... Botanist
                                           1984 hybrid species, ...
                               1897
                                           1997 confim and refin...
4 Chien-S... Physicist
                               1912
5 Katheri... Mathematician
                               1918
                                           2020 calculations of ...
6 Rosalin... Chemist
                               1920
                                           1958 <NA>
                               1928
                                           2016 existence of dar...
7 Vera Ru... Astronomer
8 Gladys ... Mathematician
                               1930
                                             NA mathematical mod...
                                             NA first scientist ...
9 Flossie... Virologist a...
                               1947
10 Jennife... Biochemist
                                1964
                                             NA one of the prima...
```

Praktikum 10 - Daten zusammenfügen

2er Teams

- 1. **E-Mail**: Öffne deine Email und klicke auf den Link zu deinem persönlichen GitHub repo
- 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"

Inputs - Drei Dataframes

einkaeufe preise kundenprofile

```
# A tibble: 9 x 4
 kunden_id produkt_name einkauf einheit
 <chr> <chr> <chr>
1 k1 Chips
                    2 anzahl
                      3 anzahl
2 k1
         Milch
                   1 anzahl
        Avocado
3 k1
4 k2 Pfirsich
                      2.5 kg
5 k2
         Birne
                      0.5 kg
6 k2
         Apfel
                      2 kg
7 k2
         Tomate
                      1.5 kg
8 k2
         Pfirsich
                      1 kg
9 k2
         Milch
                      4 anzahl
```

Gewünschter Output

vorname	nachname	summe	email
Edwin	Dumont	9.2	edwin.dumont@example.com
Leonora	Garcia	24.6	leonora.garcia@example.com

Schritt 1 - Daten zusammenfügen

```
einkaeufe_preise <- einkaeufe %>%
  left_join(preise)
einkaeufe_preise
```

```
# A tibble: 9 x 5
 kunden_id produkt_name einkauf einheit preis
 <chr> <chr> <dbl> <chr> <dbl>
1 k1 Chips 2 anzahl 3.8
       Milch 3 anzahl 2.2
2 k1
3 k1 Avocado 1 anzahl 3.2
                  2.5 kg 6.5
4 k2 Pfirsich
5 k2 Birne
                  0.5 kg 2.6
6 k2
       Apfel
                  2 kg 4.1
            1.5 kg 2.7
7 k2
       Tomate
8 k2
       Pfirsich
                  1 kg 6.5
                  4 anzahl 2.2
9 k2
       Milch
```

Schritt 2 - Daten zusammefassen

```
einkaeufe_preise_sum <- einkaeufe_preise %>%
  group_by(kunden_id) %>%
  summarise(
    summe = sum(preis)
)
einkaeufe_preise_sum
```

Schritt 3 - Daten zusammenfügen + eingrenzen

```
kunden_tab <- einkaeufe_preise_sum %>%
  left_join(kundenprofile) %>%
  select(ends_with("name"), summe, email)
kunden_tab
```

Schritt 4 - Daten als Tabelle darstellen

kunden_tab %>%
 gt()

vorname	nachname	summe	email
Edwin	Dumont	9.2	edwin.dumont@example.com
Leonora	Garcia	24.6	leonora.garcia@example.com

Als eine Code Sequenz

```
einkaeufe %>%
  left_join(preise) %>%
  group_by(kunden_id) %>%
  summarise(
    summe = sum(preis)
) %>%
  left_join(kundenprofile) %>%
  select(ends_with("name"), summe, email) %>%
  gt()
```

vorname	nachname	summe	email
Edwin	Dumont	9.2	edwin.dumont@example.com
Leonora	Garcia	24.6	leonora.garcia@example.com

Funktionen

Funktionen - Bonusmaterial

Praktikum 11 - Funktionen

Live Coding

- 1. **E-Mail**: Öffne deine Email und klicke auf den Link zu deinem persönlichen GitHub repo
- 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"

The End

The End - Noch nicht ganz

Was habt ihr gelernt?

- Anwendung von Tidyverse Packages zum
 - o Importieren,
 - Aufräumen (Tidying),
 - Transformieren,
 - Visualisieren, und
 - Kommunizieren von Daten.
- Kollaboration und Versionsverwaltung mit Git/GitHub
- Datenprojekte reproduzierbar publizieren mit GitHub
- Das Konzept von Tidy Data

Wie geht's weiter?

Raus aus der RStudio Cloud

- 1. Installationen: https://github.com/rstatsZH/kochbuch/tree/main/01-Installation
- 2. Einmalig: Tidyverse Packages installieren
- 3. Danach: Tidyverse Packages laden
- 4. Packages ausserhalb des Tidyverse installieren und laden (e.g. janitor)

```
# Einmalig in Konsole ausführen
install.packages("tidyverse")

# In jedem Skript
library(tidyverse)
```

Weiterführende Ressourcen - Üben + Vertiefen

https://rstatszh.github.io/website/posts/2021-04-30-woche07/

Projektarbeit - Unterstützung bis Anfang Juli

Hausaufgabe 6

- 1. GitHub Repository erstellen und RStudio Projekt aufgleisen (Hausaufgabe 6)
- 2. Daten für das Projekt identifizieren

Wie es weiter geht: Bericht mit R Markdown schreiben

- 1. Daten importieren
- 2. Daten (visuell) erkunden
- 3. Daten ggf. transformieren und dann erneut (visuell) erkunden
- 4. Fragen an den Datensatz formulieren
- 5. Versuchen zu Antworten zu kommen und dokumentieren
- 6. Immer wieder, git add, commit, push

Reflexion

Reflexion

5 min Nachdenken + Notizen, dann Austausch 2er Teams

- 1. Was sind die drei nützlichsten Dinge die du gelernt hast?
- 2. Welches Thema war besonders schwer zu folgen?
- 3. Was hat dir gefehlt?

Magst du mir ein Kommentar hinterlassen?

Long answer text

Feedback

Ziele erreicht?

Bitte ausfüllen: kutt.it/rstatszh-eval



Photo by: Virgil Cayasa

Wie es für mich weiter geht

- 1. Beratung: Projektbezogener Support, Code Review, Coaching
- 2. rstatsZH Kursleitung: Info über den Kurs verbreiten
- 3. Kurse zu vertiefenden Themen: Entwicklung von 4-Stunden Workshops

Contact: Lars@Lse.de



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_d07-data-join/e1_d07-data-join.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 lizensiert sind.