Daten Transformation mit dplyr

rstatsZH - Data Science mit R

Lars Schöbitz

Oct 8, 2024

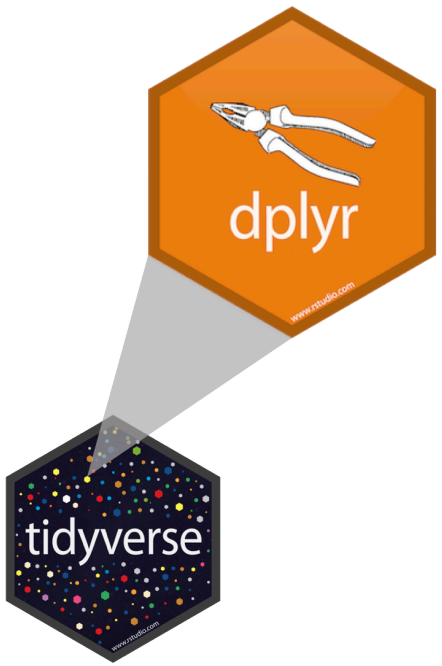
Lernziele (für diese Woche)

1. Die Lernenden können fünf Funktionen aus dem R-Paket dplyr anwenden, um eine Teilmenge von Daten zur Verwendung in einer Tabelle oder einem Diagramm zu erzeugen.
2. Die Lernenden können Funktionen aus dem R-Paket dplyr anwenden, um Daten mittels deskriptiver Statistik zusammenzufassen.

Datentransformation mit dplyr

Eine Grammatik der Datenmanipulation...

... basierend auf den Konzepten von Funktionen als Verben, die Dataframes manipulieren



- `filter`: wählt Zeilen aus, die den Kriterien entsprechen
- `arrange`: Zeilen neu ordnen
- `select`: Spalten nach Namen auswählen
- `rename`: Spalten umbenennen
- `mutate`: neue Variablen hinzufügen
- `summarise`: Variablen auf Werte reduzieren
- `group_by`: für gruppierte Operationen
- ... (viele mehr)

dplyr rules

Regeln der `dplyr`-Funktionen:

- Das erste Argument ist immer ein Dataframe.
- Nachfolgende Argumente sagen, was mit diesem Dataframe geschehen soll.
- Gibt immer einen Dataframe zurück.
- Nichts wird an Ort und Stelle verändert.

Funktionen & Argumente

```
1  library(dplyr)
2
3  filter(.data = gapminder,
4         year == 2007)
```

- Funktion: `filter()`
- Argument: `.data =`
- Argumente, die folgen: `year == 2007` Was ist mit den Daten gemacht wird

Objekte

```
1  library(dplyr)
2
3  gapminder_2007 <- filter(.data = gapminder,
4                           year == 2007)
```

- Funktion: `filter()`
- Argument: `.data =`
- Argumente, die folgen: `year == 2007` Was ist mit den Daten gemacht wird
- Daten (Objekt): `gapminder_2007`

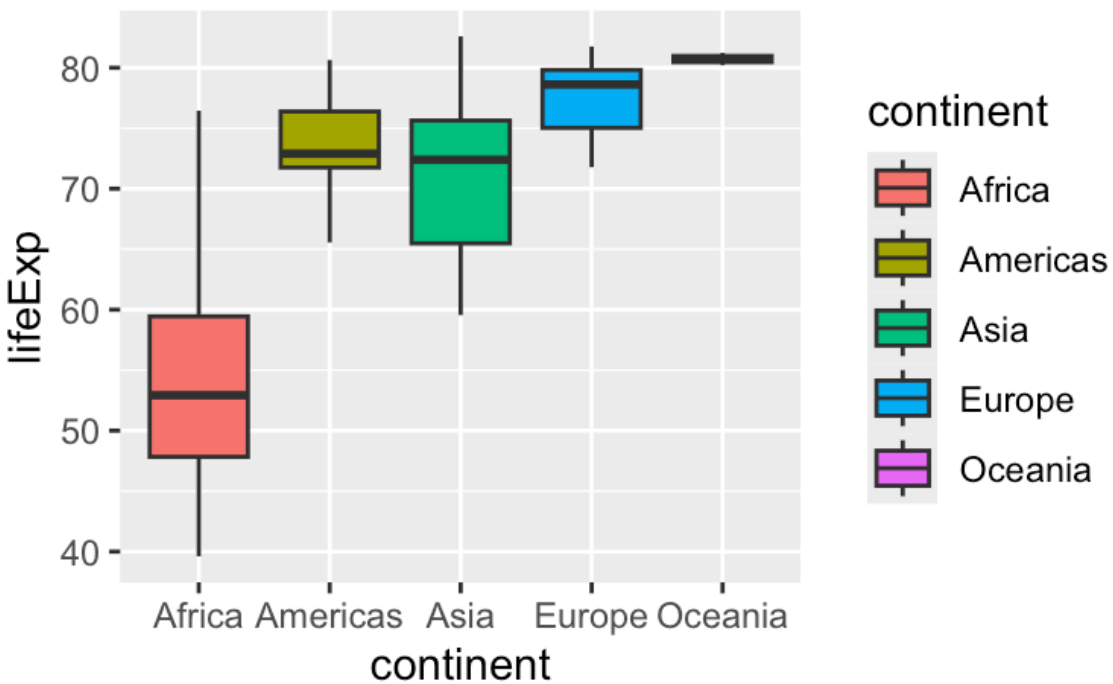
Operatoren

```
1 library(dplyr)
2
3 gapminder_2007 <- gapminder |>
4   filter(year == 2007)
```

- Funktion: `filter()`
- Argument: `.data =`
- Argumente, die folgen: `year == 2007` Was ist mit den Daten gemacht wird
- Daten (Objekt): `gapminder_2007`
- Zuweisungsoperator: `<-`
- Pipe Operator: `|>`

Grafik

```
1  library(dplyr)
2
3  gapminder_2007 <- gapminder |>
4    filter(year == 2007)
5
6  ggplot(data = gapminder_2007,
7         mapping = aes(x = continent,
8                       y = lifeExp,
9                       fill = continent)) +
10   geom_boxplot(outlier.shape = NA)
```



Wir sind dran: Treibhausgasemissionen im Kanton Zürich

rstatszh-k009.github.io/website/

Daten

```
1 treibhausgase <- read_csv("daten/ktzh-treibhausgase.csv")
```

```
1 head(treibhausgase)
```

jahr	hauptgruppe	untergruppe	thg	thg_agg	emission
1990	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0
1991	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0
1992	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0
1993	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0
1994	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0
1995	Abwasser und Abfall	Abfalldeponie	CO2	CO2eq	0

```
1 ncol(treibhausgase)
```

```
[1] 6
```

```
1 nrow(treibhausgase)
```

```
[1] 1980
```

Data

```
1 treibhausgase |>
2   distinct(hauptgruppe, untergruppe)
```

hauptgruppe	untergruppe
Abwasser und Abfall	Abfalldeponie
Abwasser und Abfall	Abwasserbehandlung
Abwasser und Abfall	Abfallverbrennung
Landwirtschaft	Fermentation bei der Verdauung
Landwirtschaft	Wirtschaftsdünger-Management
Landwirtschaft	Landwirtschaftliche Böden
Verkehr	Motorräder
Verkehr	Personenwagen
Verkehr	Linien-/Omnibusse
Verkehr	Reisebusse
Verkehr	Lastkraftwagen
Verkehr	Sattelzugmaschinen

hauptgruppe	untergruppe
Landwirtschaft	Land- und forstwirtschaftliche Maschinen
Industrie	Industrielle Fahrzeuge und Baumaschinen
Verkehr	Schiene
Verkehr	Schiff
Gebäude	Heizkessel Dienstleistungen Heizöl
Gebäude	Heizkessel Haushalte Heizöl
Gebäude	Heizkessel Dienstleistungen Gas
Gebäude	Heizkessel Haushalte Gas
Abwasser und Abfall	KVA
Industrie	Industrielle Prozesse nichtenergetisch
Industrie	Industrielle Prozesse energetisch

Wir sind dran: md-03-uebungen

1. Öffne posit.cloud in deinem Browser (verwende dein Lesezeichen).
2. Öffne den rstatszh-k009 Arbeitsbereich (Workspace) für den Kurs.
3. Klicke auf `Start` neben `md-03-uebungen`.
4. Suche im Dateimanager im Fenster unten rechts die Datei `01-dplyr-wir.qmd` und klicke darauf, um sie im Fenster oben links zu öffnen.

Pause machen

Bitte steh auf und beweg dich. Lasst eure E-Mails in Frieden ruhen.

Ihr seid dran: 02-dplyr-ihr.qmd

1. Öffne posit.cloud in deinem Browser (verwende dein Lesezeichen).
2. Öffne den rstatszh-k009 Arbeitsbereich (Workspace) für den Kurs.
3. Klicke auf `Continue` neben `md-03-uebungen`.
4. Suche im Dateimanager im Fenster unten rechts die Datei `02-dplyr-ihr.qmd` und klicke darauf, um sie im Fenster oben links zu öffnen.
5. Folge den Anweisungen in der Datei.

R Terminologie

```r
library(dplyr)

treibhausgase_verkehr <- treibhausgase |>
    filter(hauptgruppe == "Verkehr", jahr == 2022)
```

- Funktion: `filter()`
- Argumente, die folgen: `hauptgruppe == "Verkehr"` Was ist mit den Daten gemacht wird
- Daten (Objekt): `treibhausgase_verkehr`
- Zuweisungsoperator: `<-`
- Pipe Operator: `|>`

Aufgabe: Verbundenes Streudiagram

👍 Großartig für Zeitreihendaten 📅

1. Nutze die Daten treibhausgase_gebaeude und die Funktion `ggplot()`, um ein verbundenmes Streudiagramm mit `geom_point()` und `geom_line()` zu erstellen

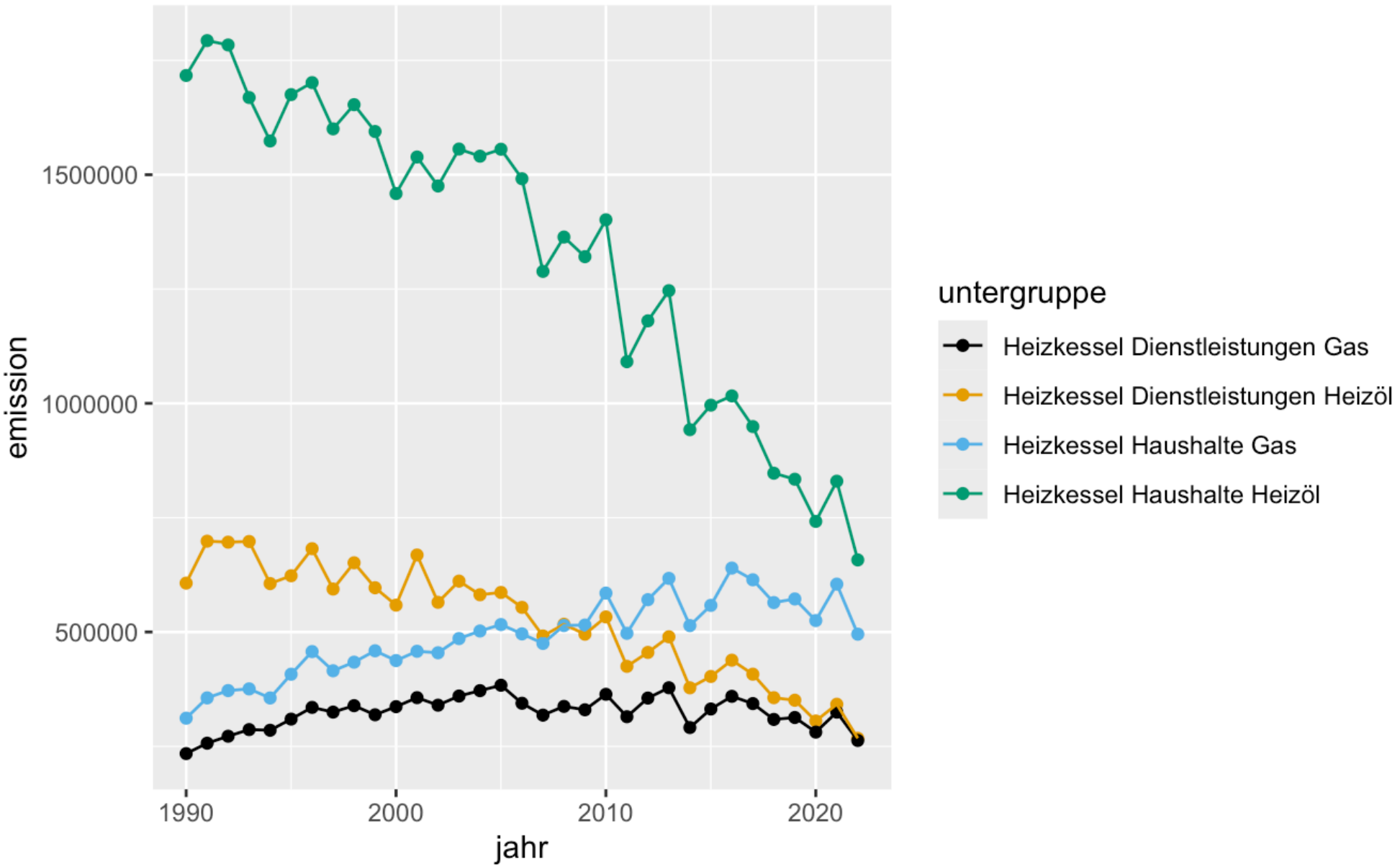
Definiere folgende visuellen Eigenschaften:

- jahr auf der x-Achse;
- emission auf der y-Achse;
- untergruppe zur Einfärbung mit dem Argument `color = untergruppe` innerhalb von `aes()`

3. Ändere die Farben mit `scale_color_colorblind()`.

```
treibhausgase |>
    filter(hauptgruppe == "Gebäude") |>
    ggplot(aes(x = jahr, y = emission, color = untergruppe)) +
    geom_point() +
```

Aufgabe: Verbundenes Streudiagram



rstatszh-k009.github.io/website/

Wir sind dran: 03-dplyr-wir.qmd

1. Öffne posit.cloud in deinem Browser (verwende dein Lesezeichen).
2. Öffne den rstatszh-k009 Arbeitsbereich (Workspace) für den Kurs.
3. Klicke auf `Start` neben `md-03-uebungen`.
4. Suche im Dateimanager im Fenster unten rechts die Datei `03-dplyr-wir.qmd` und klicke darauf, um sie im Fenster oben links zu öffnen.

Pause machen

Bitte steh auf und beweg dich. Lasst eure E-Mails in Frieden ruhen.

Zeitpuffer: 03-dplyr-wir.qmd

Welche Konzepte kann ich nochmals erklären?

Zusatzaufgaben Modul 3

Modul 3 Dokumentation

rstatszh-k009.github.io/website/module/md-03.html

Zusatzaufgaben Abgabedatum

- Abgabedatum: Montag, 14. Oktober
- Korrektur- und Feedbackphase bis zu: Donnerstag, 17. Oktober

Danke

rstatszh-k009.github.io/website/

Danke! 🌻

Folien erstellt mit revealjs und Quarto: https://quarto.org/docs/presentations/revealjs/

Access slides als PDF auf GitHub