

Mit Text Daten arbeiten & KTZH Corporate Design mit statR

rstatsZH - Data Science mit R

Lars Schöbitz

Apr 15, 2025

Lernziele (für diese Woche)

1. Die Lernenden können die Funktion `str_detect()` aus dem R-Paket `stringr` verwenden um das Auftreten oder Fehlen bestimmter Muster in Zeichenvektoren (character Vektor) zu ermitteln.
2. Die Lernenden können `str_detect()` mit `dplyr` Funktionen wie `filter()` oder `mutate()` nutzen, um Daten über das Auftreten von Mustern in Teilmengen zu unterteilen oder darauf basierend neue Variablen zu erstellen.
3. Die Lernenden können das `statR` R-Paket nutzen um eine Visualisierung im Corporate Design des Kanton Zürich zu erstellen.

Arbeiten mit Strings in R

- Strings -> Zeichenkette (eine folge von Zeichen)
- Werden verwendet um Textdaten darzustellen
- Können beliebige Länge haben
- Erstellt mit einfachen oder doppelten Anführungszeichen
- Sonderzeichen können mit dem Backslash \ “ausgenommen” werden

Anführungszeichen

- Erstellt mit einfachen oder doppelten Anführungszeichen

```
1 string1 <- "Dies ist eine Zeichenkette"  
2 string2 <- 'Wenn ich ein "Anführungszeichen" in eine Zeichenkette einfügen  
3           möchte, verwende ich einfache Anführungszeichen'
```

Der Backslash \

Um ein einfaches oder doppeltes Anführungszeichen in einer Zeichenkette zu verwenden, kann `\`, um es “auszunehmen”:

```
1 double_quote <- "\""  
2 single_quote <- '\''
```

Falls du ein wörtliches Backslash in deiner Zeichenkette verwenden möchtest, musst du es “ausnehmen”: `"\\`”:

```
1 backslash <- "\\"
```

Beachte dass die gedruckte Darstellung einer Zeichenkette in der Console nicht identisch mit der Zeichenkette selbst ist:

```
1 x <- c(single_quote, double_quote, backslash)  
2 x
```

```
[1] "' " "\" "\\ "
```

Um den Rohinhalt der Zeichenkette zu sehen, verwende `str_view()`

```
1 str_view(x)
```

```
[1] | '  
[2] | "  
[3] | \
```

Vornamen Statistik

Daten: Vornamen der Bevölkerung nach Jahrgang, Schweiz, 2023

- jährlich aktualisierte Daten
- Vornamen mit weniger als 3 Nennungen werden ausgeschlossen
- Datenquelle: Bundesamt für Statistik

Frage: Wieviele einzigartige Vornamen gibt es in der Schweiz?

1. 200'000
2. 1'000
3. 1.0 mio
4. 50'000

Vornamen Statistik

Frage: Wieviele einzigartige Vornamen gibt es in der Schweiz?

```
1 vornamen
```

```
# A tibble: 976,068 × 4
  vorname geburtsjahr wert geschlecht
  <chr>      <dbl> <dbl> <chr>
1 Olivier    1915     1 m
2 Florian    1917     1 m
3 Max        1917     1 m
4 Albert     1918     1 m
5 Co         1918     1 m
6 Julian     1918     1 m
7 Victor     1918     1 m
8 Alfred     1919     1 m
9 Valentin   1919     1 m
10 Walter    1919     1 m
# i 976,058 more rows
```

```
1 vornamen |>
2   distinct(vorname)
```

```
# A tibble: 65,401 × 1
  vorname
  <chr>
1 Olivier
2 Florian
3 Max
4 Albert
5 Co
6 Julian
7 Victor
8 Alfred
9 Valentin
10 Walter
# i 65,391 more rows
```

Vornamen Statistik

Frage: Was sind die häufigsten 10 Vornamen in der Schweiz?

```
1 vornamen |>
2   count(vorname, geschlecht,
3         wt = wert, sort = TRUE) |>
4   head(n = 10)
```

```
# A tibble: 10 × 3
  vorname  geschlecht    n
  <chr>    <chr>  <dbl>
1 Maria    w      74840
2 Daniel   m      62884
3 Peter    m      54007
4 Thomas   m      52732
5 Hans     m      44073
6 Christian m      41702
7 Martin   m      40627
8 Anna     w      40387
9 Michael  m      39922
10 Andreas m      39583
```


Vornamen Statistik

Frage: Was sind die häufigsten 10 Vornamen in der Schweiz?

```
1 vornamen |>
2   count(vorname, geschlecht,
3         wt = wert, sort = TRUE) |>
4   head(n = 10) |>
5   # aus dem knitr R-Paket
6   kable()
```

vorname	geschlecht	n
Maria	w	74840
Daniel	m	62884
Peter	m	54007
Thomas	m	52732
Hans	m	44073
Christian	m	41702
Martin	m	40627
Anna	w	40387
Michael	m	39922
Andreas	m	39583

Vornamen Statistik

Frage: Was sind die häufigsten 10 Vornamen in der Schweiz?

```
1 vornamen |>
2   count(vorname, geschlecht,
3         wt = wert, sort = TRUE) |>
4   head(n = 10) |>
5   # aus dem gt R-Paket
6   gt()
```

vorname	geschlecht	n
Maria	w	74840
Daniel	m	62884
Peter	m	54007
Thomas	m	52732
Hans	m	44073
Christian	m	41702
Martin	m	40627
Anna	w	40387
Michael	m	39922
Andreas	m	39583

Vornamen Statistik

Frage: Was sind die häufigsten 10 Vornamen in der Schweiz?

```
1 vornamen |>
2   count(vorname, geschlecht,
3         wt = wert, sort = TRUE) |>
4   head(n = 10) |>
5   # nutze gt R-Package für die Darstellung
6   gt() |>
7   tab_style(
8     style = cell_fill(color = "#AFF0ED"),
9     locations = cells_body(
10       columns = everything(),
11       rows = geschlecht == "m"
12     )
13   ) |>
14   tab_style(
15     style = cell_fill(color = "#FFD700"),
16     locations = cells_body(
17       columns = everything(),
18       rows = geschlecht == "w"
19     )
20   )
```

vorname	geschlecht	n
Maria	w	74840
Daniel	m	62884
Peter	m	54007
Thomas	m	52732
Hans	m	44073
Christian	m	41702
Martin	m	40627
Anna	w	40387
Michael	m	39922
Andreas	m	39583

Vornamen Statistik

Frage: Was ist die Verteilung der Vornamenlängen in der Schweiz?

```
1 vorkamen |>
2   count(vorname, wt = wert) |>
3   mutate(laenge = str_length(vorname)) |>
4   count(laenge) |>
5   print(n = 19)
```

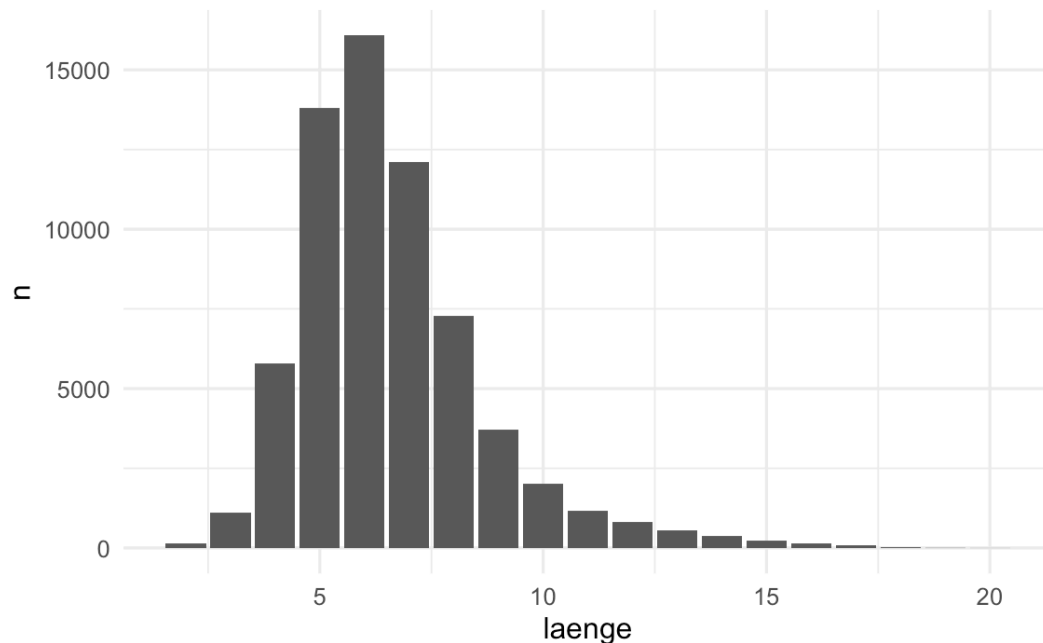
A tibble: 19 × 2

	laenge	n
	<int>	<int>
1	2	128
2	3	1094
3	4	5780
4	5	13793
5	6	16075
6	7	12114
7	8	7294
8	9	3719
9	10	2013
10	11	1155
11	12	811
12	13	559
13	14	384
14	15	233
15	16	131
16	17	75

Vornamen Statistik

Frage: Was ist die Verteilung der Vornamenlängen in der Schweiz?

```
1 vornamen_laenge_sum <- vornamen |>
2   count(vorname, wt = wert) |>
3   mutate(laenge = str_length(vorname)) |>
4   count(laenge)
5
6 ggplot(data = vornamen_laenge_sum,
7       mapping = aes(x = laenge, y = n)) +
8   geom_col() +
9   theme_minimal(base_size = 12)
```



Ihr seid dran: Vornamen Statistik

Frage: Welche Fragen könnten wir noch zu den Vornamen in der Schweiz stellen?

1. Macht ein paar Notizen.
2. Teilt sie im Chat.

stringr: Zeichenkettenmanipulation in R

Hauptmerkmale:

- Teil der tidyverse R-Pakete
- Konsistente Syntax mit str_-Präfix

Funktionen:

- `str_length()`: Stringlänge ermitteln
- `str_c()`: Strings verketteten
- `str_sub()`: Teilstrings extrahieren/ersetzen
- `str_detect()`: Mustererkennung
- `str_count()`: Anzahl Vorkommen eines Musters
- ...

Ich bin dran: stringr R-Paket

Zurücklehnen und
Fragen stellen!

Pause machen

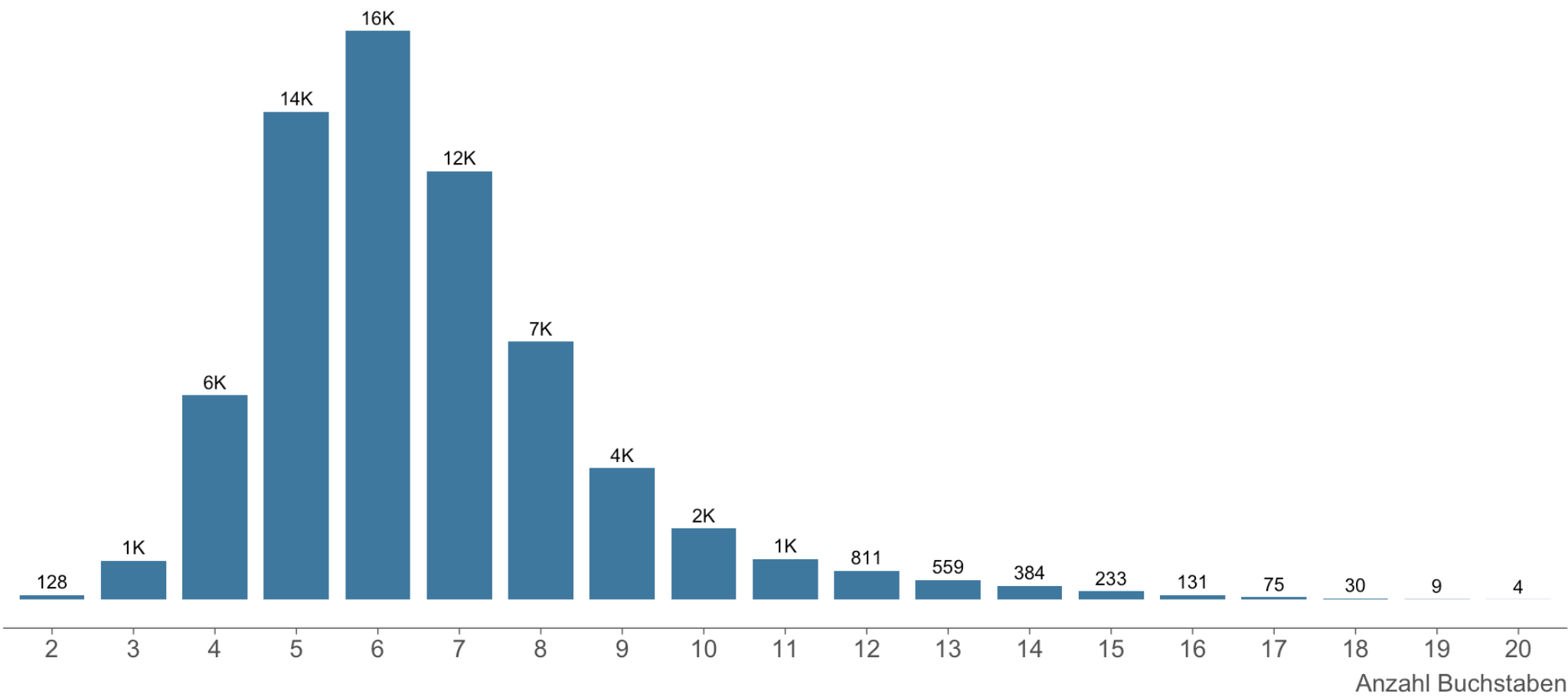
Bitte steh auf und beweg dich. Lasst eure E-Mails in Frieden ruhen.

Kanton Zürich – Corporate Design

Vornamen Statistik mit statR R-Paket

Anzahl Buchstaben in Vornamen, Schweiz

Jahrgänge 1915 bis 2023



statR R-Paket

- Erstellt Corporate Design Visualisierungen für den Kanton Zürich
- Enthält ein benutzerdefiniertes `ggplot2`-Theme
- Bietet generische Farbpaletten für Datenvisualisierungen
- Export von Datensätzen als XLSX-Dateien mit Quellinformationen und zusätzlichen Metadaten
- Stellt eine HTML-Berichtsvorlage zur Verfügung
- Offen auf GitHub verfügbar:
<https://github.com/statistikZH/statR>

Wir sind dran: 02-statR-wir.qmd

1. Öffne posit.cloud in deinem Browser (verwende dein Lesezeichen).
2. Öffne den rstatszh-k010 Arbeitsbereich (Workspace) für den Kurs.
3. Klicke auf **Start** neben **md-08-uebungen**.
4. Suche im Dateimanager im Fenster unten rechts die Datei **02-statR-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: 03-vornamen-ihr.qmd

1. Öffne posit.cloud in deinem Browser (verwende dein Lesezeichen).
2. Öffne den rststatszh-k010 Arbeitsbereich (Workspace) für den Kurs.
3. Klicke auf **Continue** neben **md-08-uebungen**.
4. Suche im Dateimanager im Fenster unten rechts die Datei **03-vornamen-ihr.qmd** und klicke darauf, um sie im Fenster oben links zu öffnen.
5. Folge den Anweisungen in der Datei.

Zeitpuffer: Modul 8

Kann ich noch etwas zum
heutigen Modul erklären?

Zusatzaufgaben Modul 8

Modul 8 Dokumentation

rstatszh-k010.github.io/website/module/md-08.html

Zusatzaufgaben Abgabedatum

- Abgabedatum: Montag, 21. April
- Korrektur- und Feedbackphase bis zu: Donnerstag, 24. April

Danke

Danke!

Folien erstellt mit revealjs und Quarto:

<https://quarto.org/docs/presentations/revealjs/>

Access slides als [PDF auf GitHub](#)

Alle Materialien sind lizenziert unter [Creative Commons Attribution Share Alike 4.0 International](#).