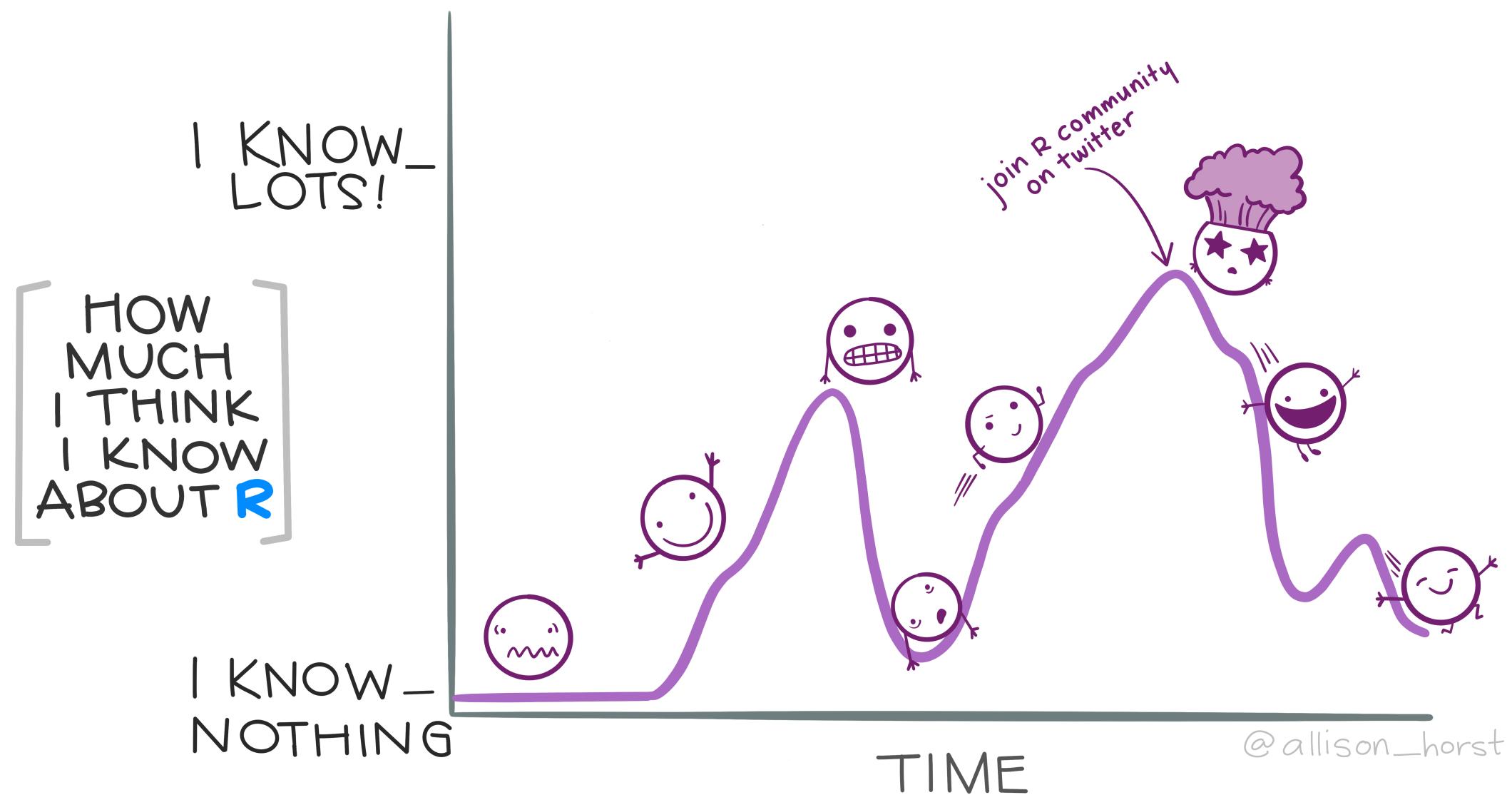


Explorative Datenanalyse durch Visualisierung & Digitaler Arbeitsplatz (DAP)

rstatsZH - Data Science mit R

Lars Schöbitz

Oct 1, 2024
 rstatszh-k009.github.io/website/



Lösung von Coding Problemen

Tipps für Suchmaschinen

- Verwende Verben, die beschreiben, was du tun willst
- Sei präzise
- Füge R zur Suchanfrage hinzu
- Füge den Namen des R-Pakets zur Suchanfrage hinzu (z.B ggplot2)
- Scrolle durch die ersten 5 Ergebnisse (wähle nicht nur das erste aus)
- Schreibe die Suchanfrage auf Englisch

Beispiel: How to remove a legend from a plot in R ggplot2?

Stack Overflow

Was ist das?

- Das größte Unterstützungsnetzwerk für (Coding-)Probleme
- Kann anfangs einschüchternd sein
- Upvote-System

Arbeitsablauf

- Lies dir zuerst kurz die Frage durch, die gepostet wurde.
- Lies dir dann die Antwort durch, die als “richtig” markiert wurde.
- Lies dir dann eine oder zwei weitere Antworten mit vielen Zustimmungen durch.
- Sieh dir dann die “linked posts” an.

Tipps für AI Werkzeuge

- Verwende Verben, die beschreiben, was du tun willst
- Präzise sein ist weniger wichtig
- Füge R zur Suchanfrage hinzu
- Füge den Namen des R-Pakets zur Suchanfrage hinzu (z.B ggplot2)
- Schreibe die Suchanfrage auf Englisch oder Deutsch

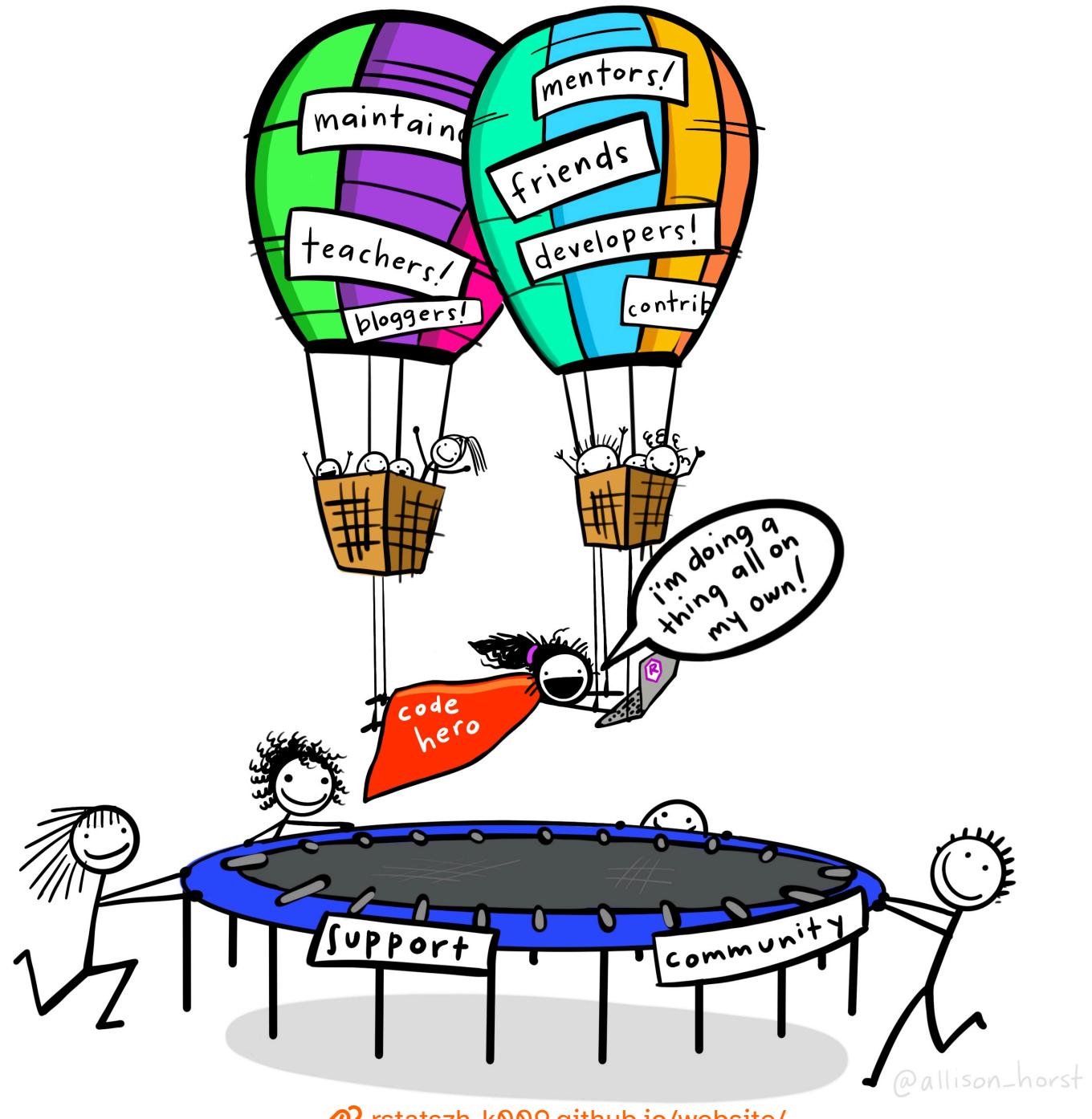
Wie entferne ich eine Legende aus einem Diagramm in R ggplot2?

Legende in R ggplot2 entfernen.

Andere Quellen für Hilfe

- Posit Community Forum:
<https://community.rstudio.com/>
- Dokumentation Webseiten:
<https://ggplot2.tidyverse.org/>
- Mastodon tag: [#rstats](#)





Lernziele (für diese Woche)

Digitaler Arbeitsplatz / R Community

Triff das STAT

Philipp Bosch

- Job ohne Excel gesucht 🗅
- Community Mensch ❤️
- Data Literacy Fan 💡

Thomas Knecht

- R-Support gegen Cookies 🍪
- R-Infrastruktur Guru 🧙
- Debug-Master & Kartentyp 🌎

Installation auf dem DAP

Um in den vollen Genuss der R-Analyse Umgebung zu kommen, musst du im Serviceportal folgende Module bestellen:

Kategorien

⊕ Arbeitsplatz

⊕ Smart Devices

⊕ Applikationen

Standardapplikationen

Berechtigungen

Fachapplikationen JI

Schriftarten (Fonts)

Applikationen bestellen

Applikationen kündigen

⊕ ServiceNow

⊕ Services

Neuer Bedarf

Servicekatalog & Preise

R for Windows

UID 2208 - R for Windows 4.3.1

R Studio

UID 2209 - R Studio ist eine integrierte Entwicklungsumgebung (IDE) für R, optimiert für Datenanalyse, Visualisierung und statistische Berechnungen.

Rtools

UID 2210 - Rtools

TinyTeX

UID 2211 - TinyTeX

Community

Im Kanton haben wir eine Community of Practice für R, welche ihre digitale Heimat in einem Teams-Kanal hat.

[Hier geht's zum Kanal](#)

Im Kanal könnt ihr:

- Fragen rund um R stellen (Stackoverflow des Kantons)
- Neue Infos zum R-Bundle erhalten
- Up-to-date bleiben was in der Community läuft

R-Fachgruppe

Aufgaben

- Updates des R-Bundles
- Weiterentwicklung der Installation anhand eurer Bedürfnisse
- Support bei Installationsproblemen

Vertretungen

- Thomas Knecht (JI)
- Philipp Bosch (JI)
- Sarah Gerhard (BI)
- Miriam Hofstetter (VD)
- Andreas Gubler (BD)
- Gianluca Macauda (GD)
- Fabian Berger (SD)
- Jörg Sintermann (BD)
- Nina Schnyder (JI)
- Gian-Marco Alt (BD)
- Joëlle Ninon Albrecht (JI)

Ihr seid dran: Fragen

Stellt eure Fragen an
Thomas und Philipp

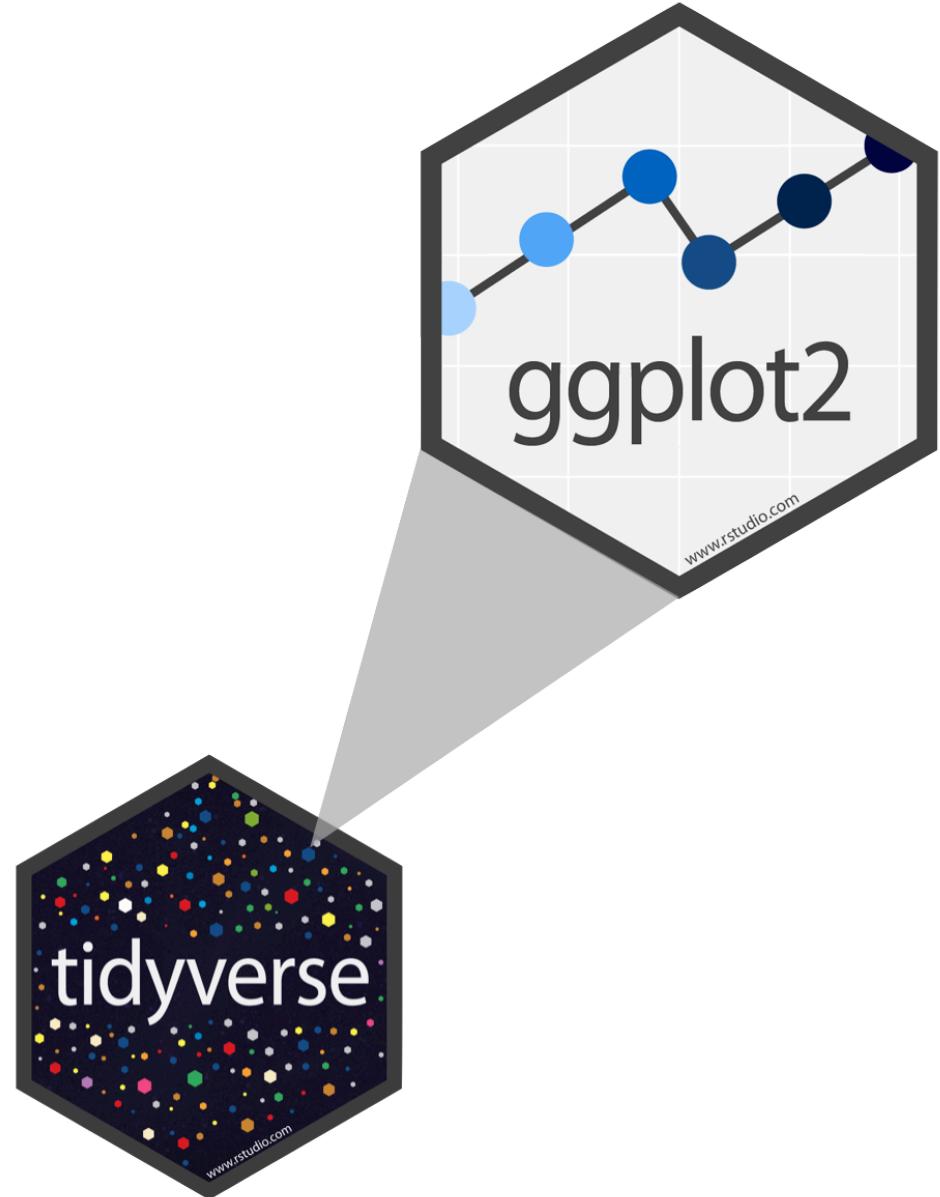
Pause machen

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

Explorative Datenanalyse mit `ggplot2`

R Paket ggplot2

- ggplot2 ist das Datenvisualisierungspaket von tidyverse
- gg“ in “ggplot2” steht für “Grammar of Graphics”
- Inspiriert durch das Buch **Grammar of Graphics** von Leland Wilkinson
- Dokumentation: <https://ggplot2.tidyverse.org/>
- Buch: <https://ggplot2-book.org>



Ich bin dran: Arbeiten mit Quarto und R

zurücklehnen und
genießen!

Code Struktur

- `ggplot()` ist die Hauptfunktion von ggplot2
- Plots werden in Schichten aufgebaut
- Die Struktur des Codes für Plots lässt sich wie folgt zusammenfassen

```
1 ggplot(data = [datensatz],  
2         mapping = aes(x = [x-variable],  
3                           y = [y-variable])) +  
4         geom_xxx() +  
5         andere Optionen
```

Code Struktur

```
1 ggplot()
```

Code Struktur

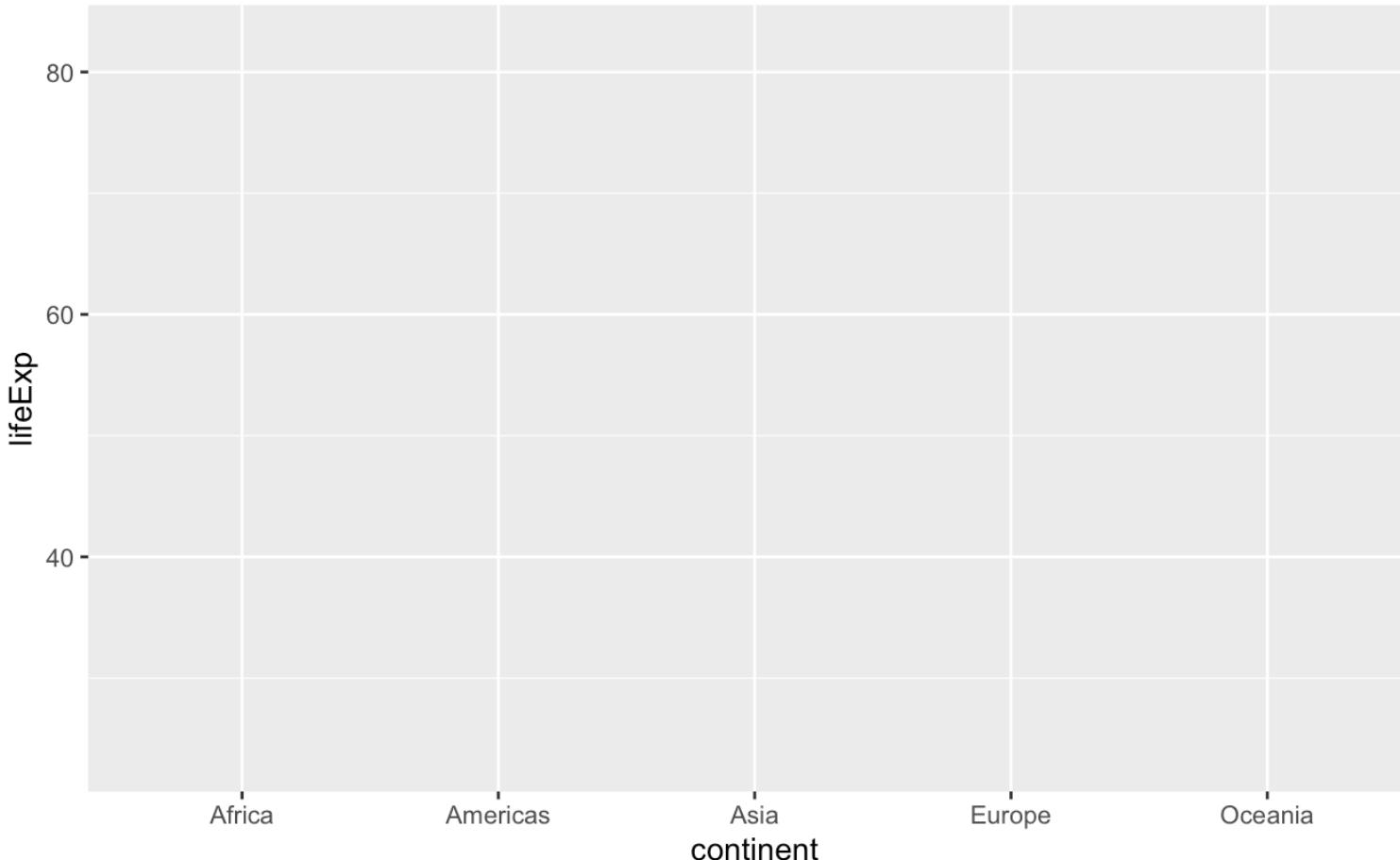
```
1 ggplot(data = gapminder)
```

Code Struktur

```
1 ggplot(data = gapminder,  
2         mapping = aes())
```

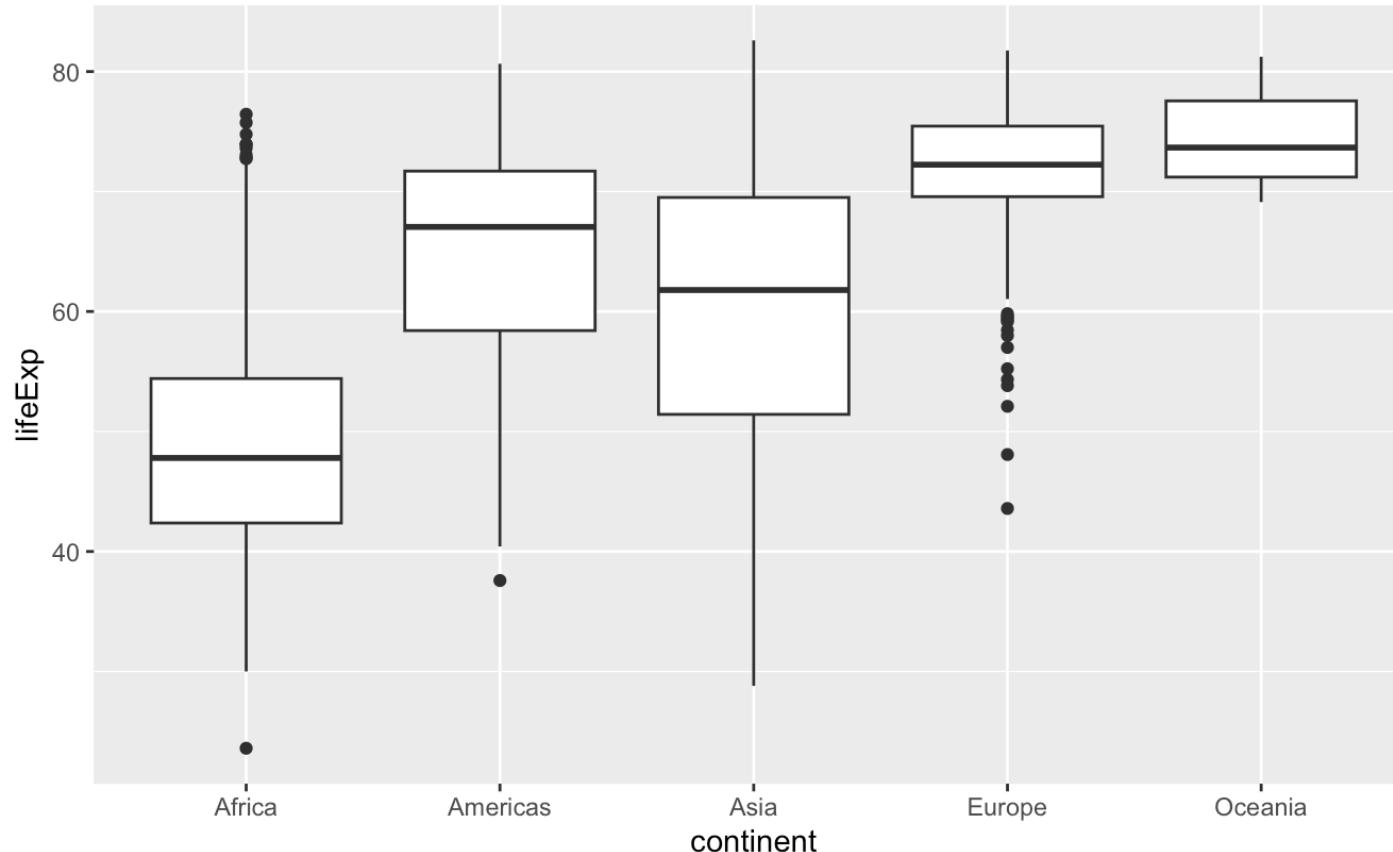
Code Struktur

```
1 ggplot(data = gapminder,  
2         mapping = aes(x = continent,  
3                           y = lifeExp))
```



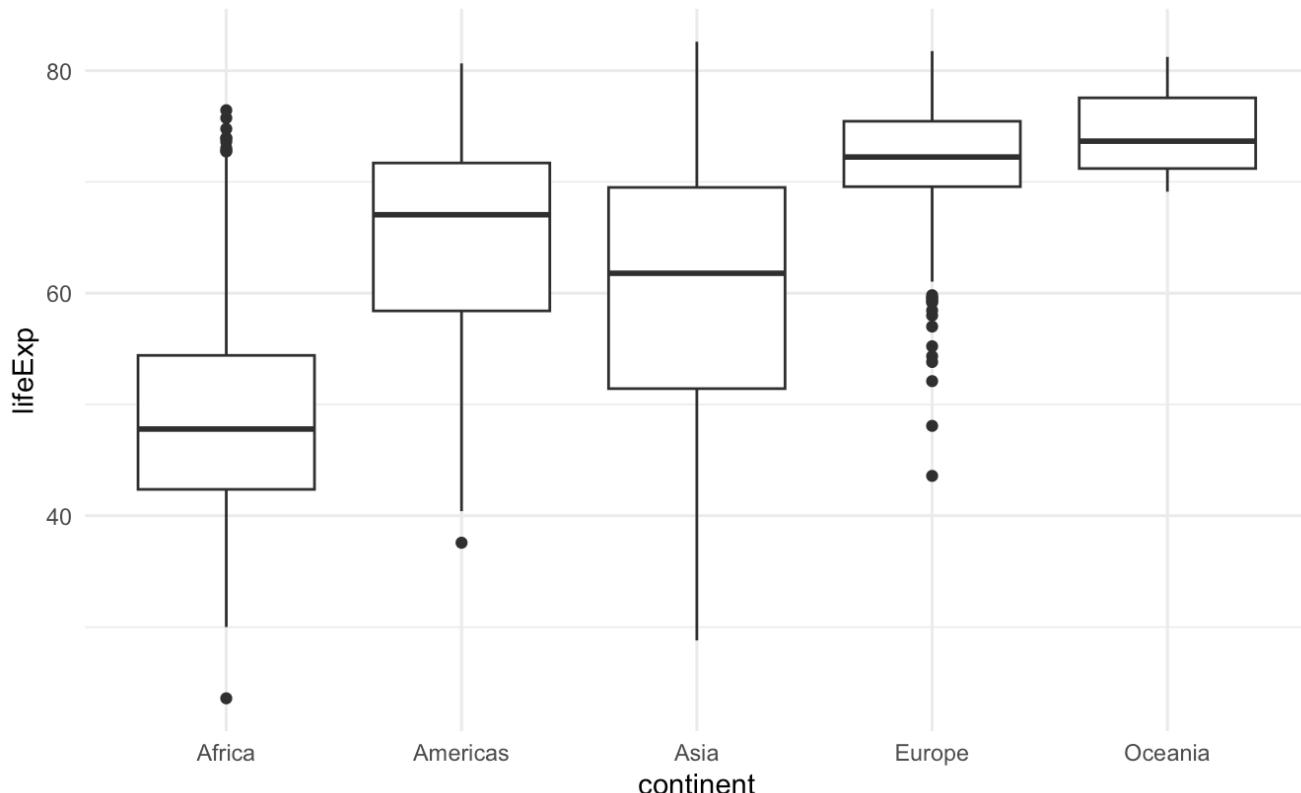
Code Struktur

```
1 ggplot(data = gapminder,  
2         mapping = aes(x = continent,  
3                           y = lifeExp)) +  
4   geom_boxplot()
```



Code Struktur

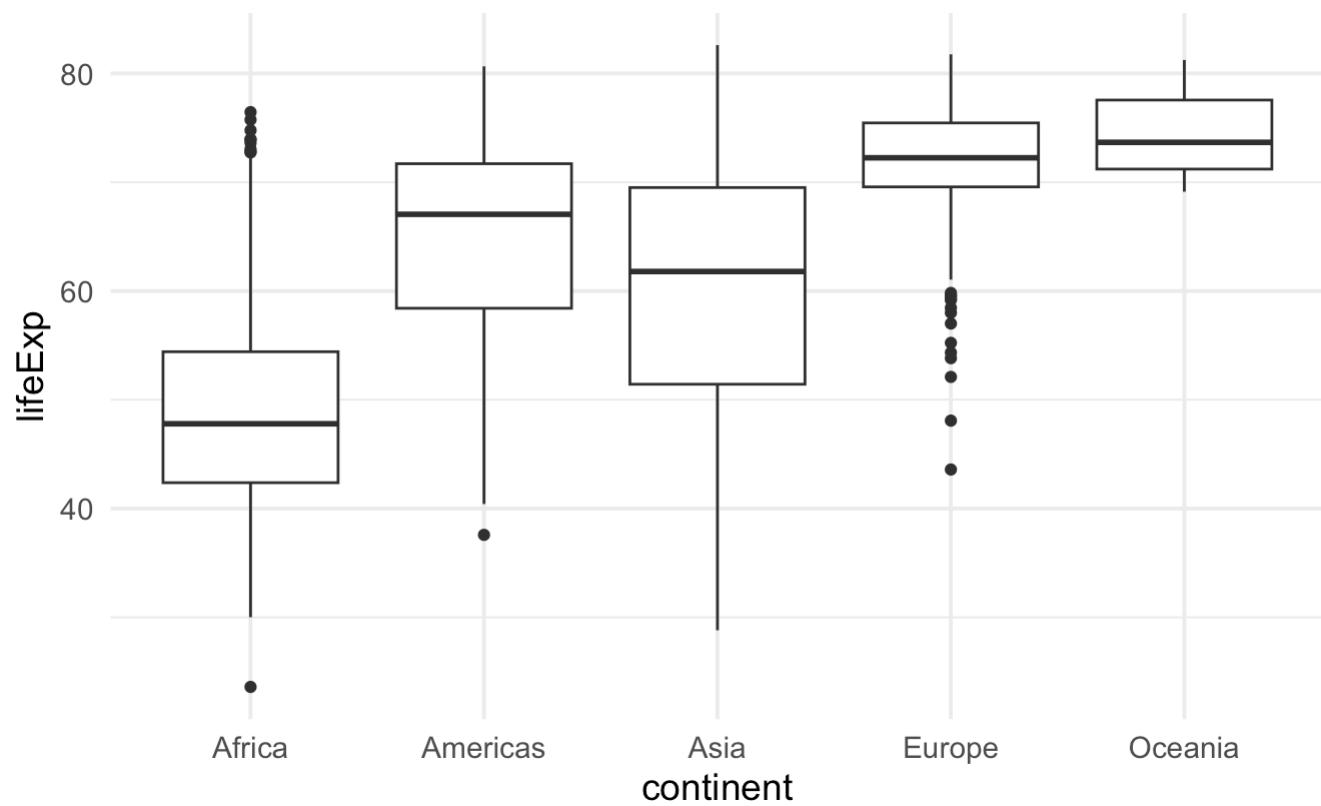
```
1 ggplot(data = gapminder,  
2         mapping = aes(x = continent,  
3                           y = lifeExp)) +  
4   geom_boxplot() +  
5   theme_minimal()
```



Polls

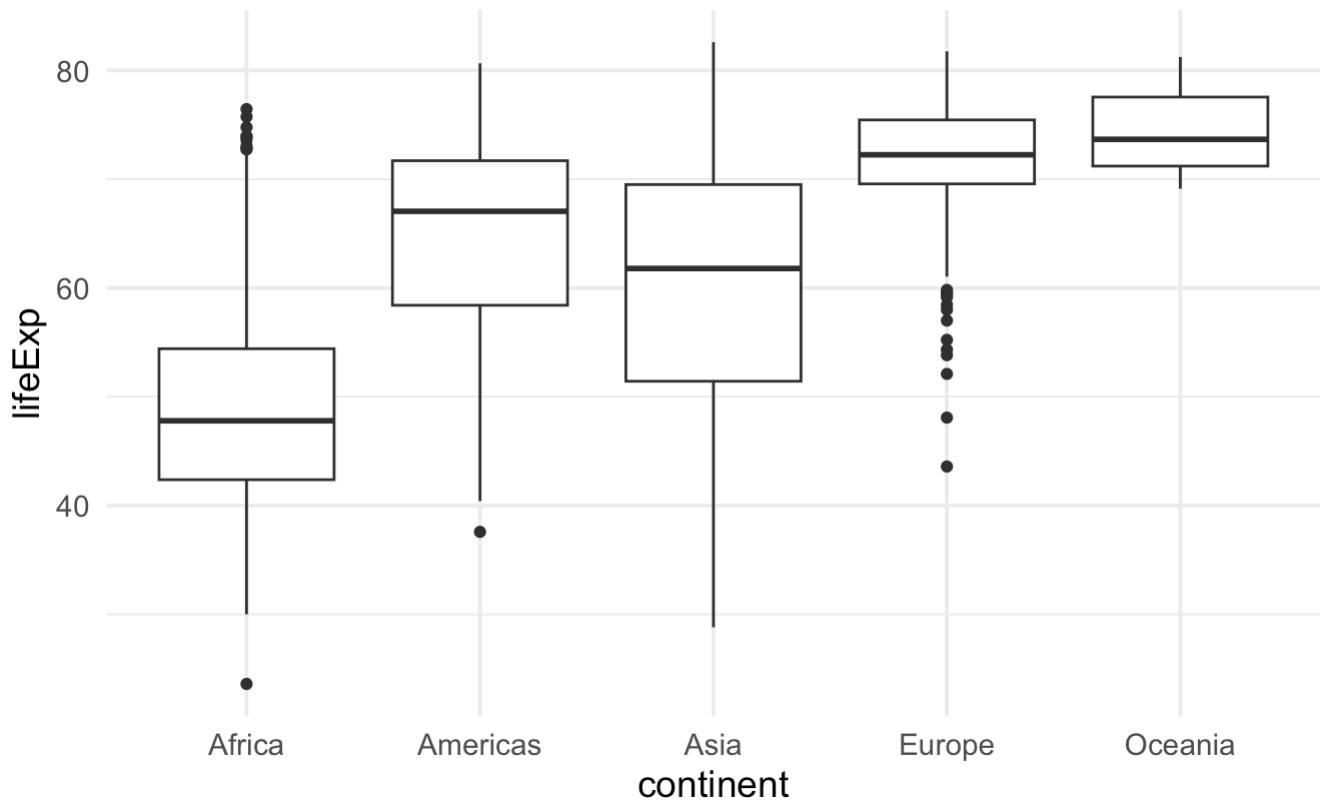
Poll 1: Was stellt die dicke Linie innerhalb des Kastens eines Boxplots dar?

1. Ich weiß es nicht
2. der Mittelwert der Beobachtungen
3. die Mitte der Box
4. der Median der Beobachtungen



Poll 2: Wie viel Prozent der Beobachtungen befinden sich innerhalb der Box eines Boxplots (Interquartilsbereich)?

1. Ich weiß es nicht
2. 25%
3. hängt vom Median ab
4. 50%



Poll 3: Was ist der Median einer Gruppe von Beobachtungen?

1. Ich weiß es nicht
2. Der Median ist der am häufigsten vorkommende Wert in einem Datensatz.
3. Der Median ist die Summe aller Werte in einem Datensatz geteilt durch die Anzahl der Beobachtungen.
4. Der Median ist der Punkt, über und unter dem die Hälfte (50%) der Beobachtungen liegt.

Boxplot, erklärt

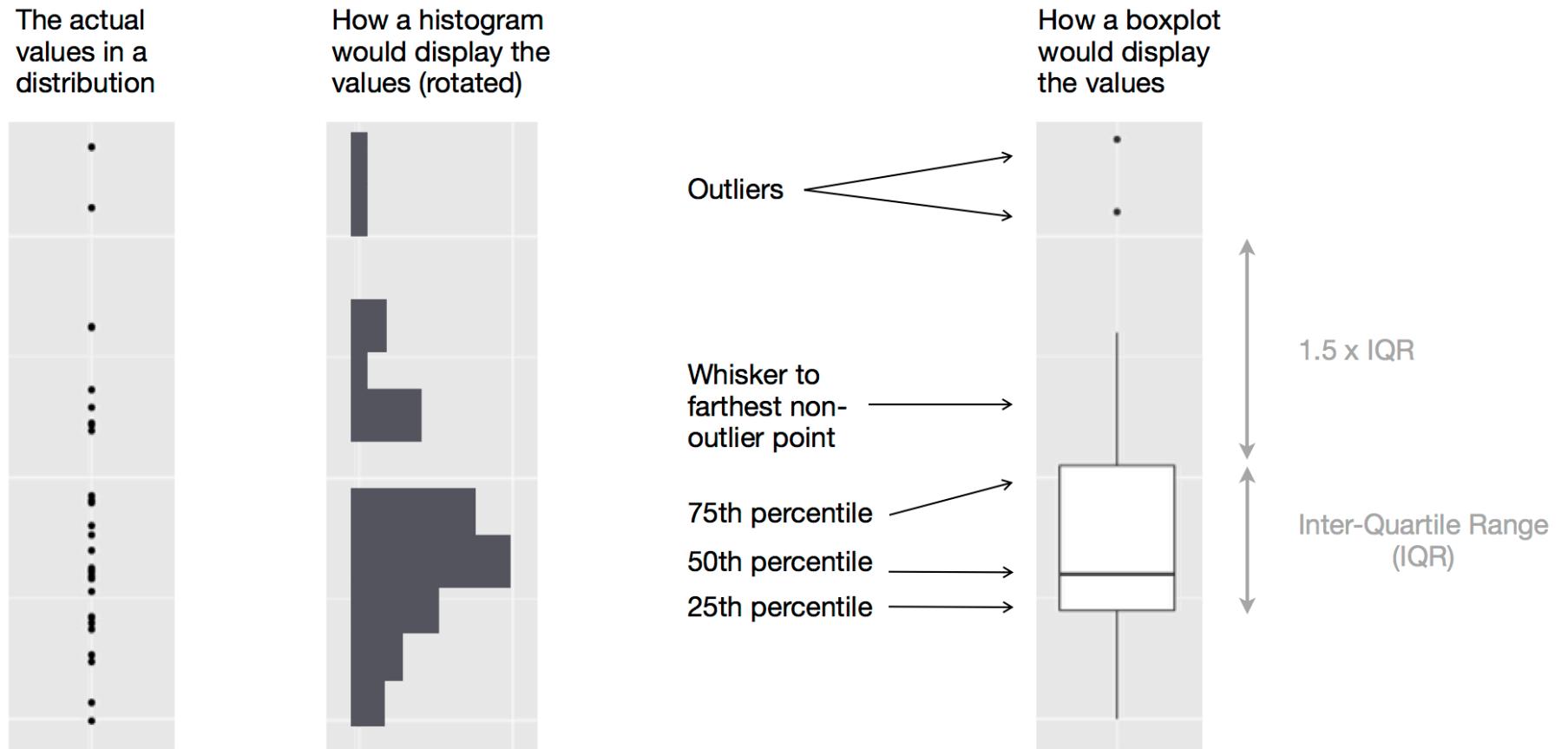


Figure 1: Diagramm, das zeigt, wie ein Boxplot erstellt wird.

Wir sind dran: md-02-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-02-uebungen.
4. Suche im Dateimanager im Fenster unten rechts die Datei md-02b-daten-visualisierung.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.

Daten visualisieren

Arten von Variablen

Numerisch

Diskrete Variablen

- nicht negative
- zählbare
- ganze Zahlen
- z.B. Anzahl Schüler, Würfelwurf

Stetige (kontinuierliche) Variablen

- unendliche Anzahl von Werten
- zwischen zwei Werten
- auch Datums/Uhrzeitwerte
- z.B. Länge, Gewicht, Größe

Nicht numerisch

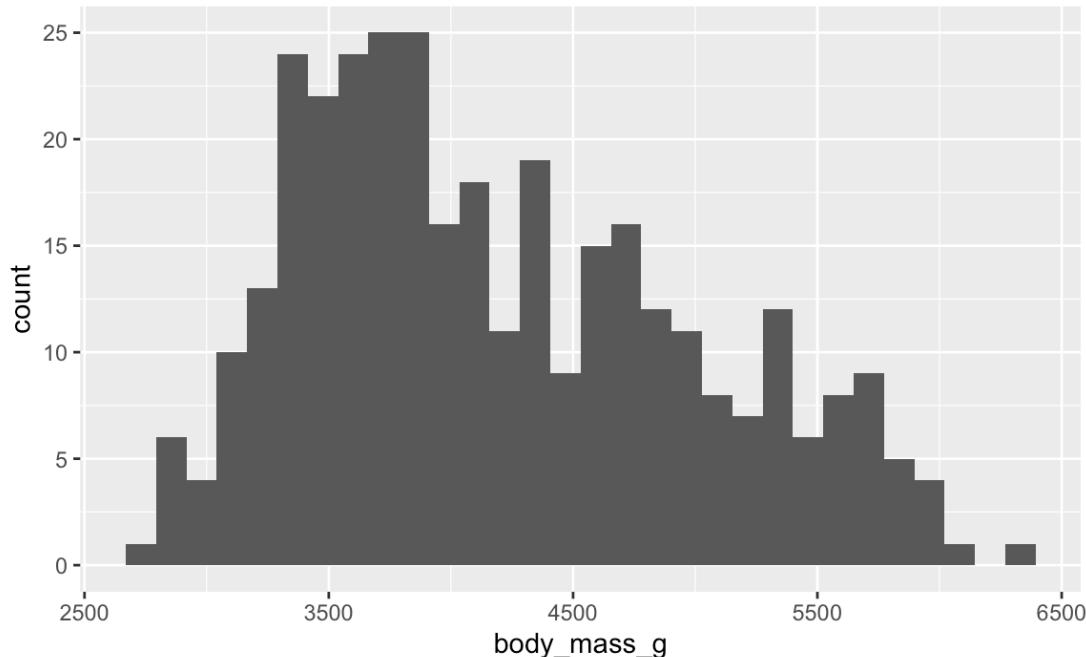
Kategoriale Variablen

- endliche Anzahl von Werten
- eindeutige Gruppen (z.B. EU Länder)
- **ordinal**, wenn diese eine logische Reihenfolge/Rangordnung aufweisen (z.B. Wochentage, Schulnoten)

Histogramm

- zur Visualisierung der Verteilung von kontinuierlichen (numerischen) Variablen

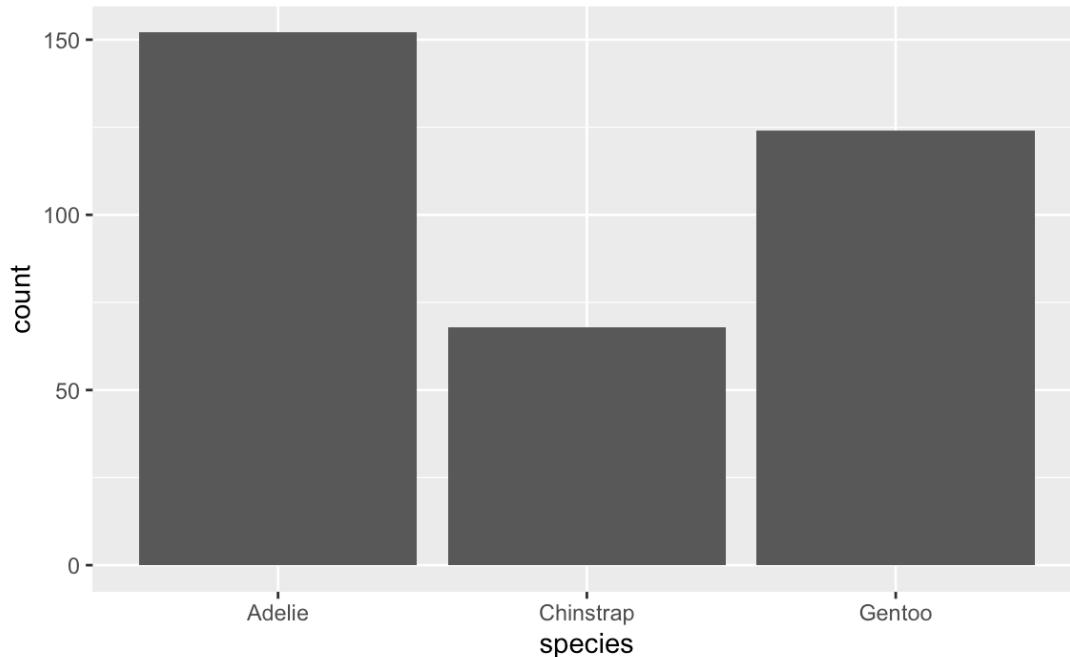
```
1 ggplot(data = penguins,
2         mapping = aes(x = body_mass_g)) +
3   geom_histogram()
```



Barplot (Säulendiagramm)

- zur Visualisierung der Verteilung von kategorischen (nicht numerischen) Variablen

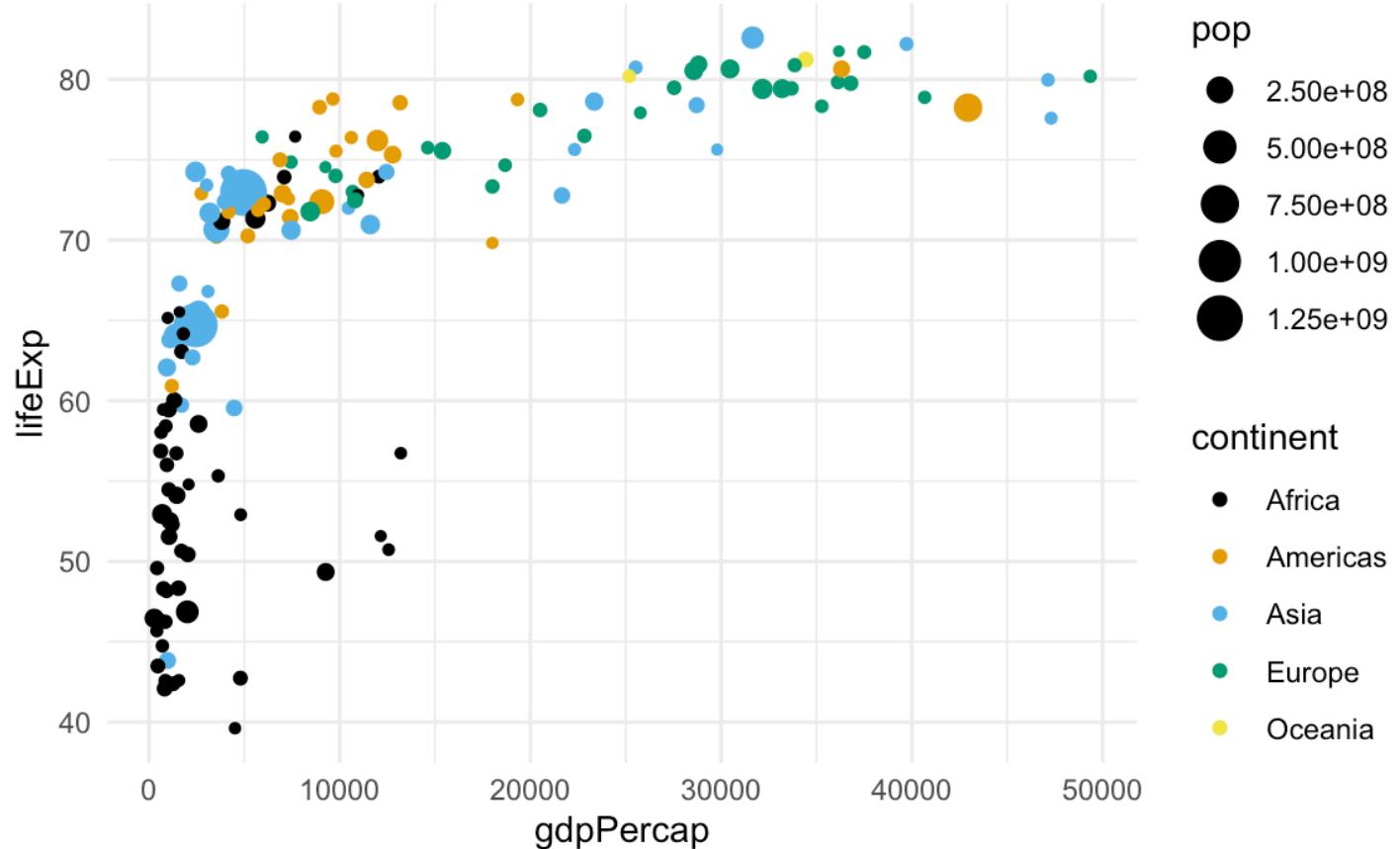
```
1 ggplot(data = penguins,
2         mapping = aes(x = species)) +
3     geom_bar()
```



Scatterplot (Streudiagramm)

- for visualizing relationships between two continuous (numerical) variables

```
1 ggplot(data = gapminder_2007,
2         mapping = aes(x = gdpPercap,
3                          y = lifeExp,
4                          size = pop,
5                          color = continent)) +
6   geom_point() +
7   scale_color_colorblind() +
8   theme_minimal()
```



Zusatzaufgaben Modul 2

Modul 2 Dokumentation

rstatszh-k009.github.io/website/module/md-02.html

Zusatzaufgaben Abgabedatum

- Abgabedatum: Montag, 07. Oktober
- Korrektur- und Feedbackphase bis zu: Donnerstag, 10. Oktober

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.