

Datenvisualisierung 3

Darstellung der zusammenfassenden Statistik

Daniela Palleschi

Humboldt-Universität zu Berlin

2023-12-18

Lernziele

Heute werden wir lernen...

- Boxplots zu erstellen und zu interpretieren
- Mittelwerte und Standardabweichungen zu visualisieren

Ressourcen

- [Abschnitt 2.5 \(Visualisierung von Beziehungen\)](#) in Wickham et al. (2023)
- [Kapitel 4 \(Darstellung von zusammenfassenden Statistiken\)](#) in Nordmann et al. (2022)
- Abschnitte 3.5-3.9 in Winter (2019)

Einrichten

Pakete

```
1 pacman::p_load(tidyverse,  
2               here,  
3               janitor,  
4               ggthemes,  
5               patchwork)
```

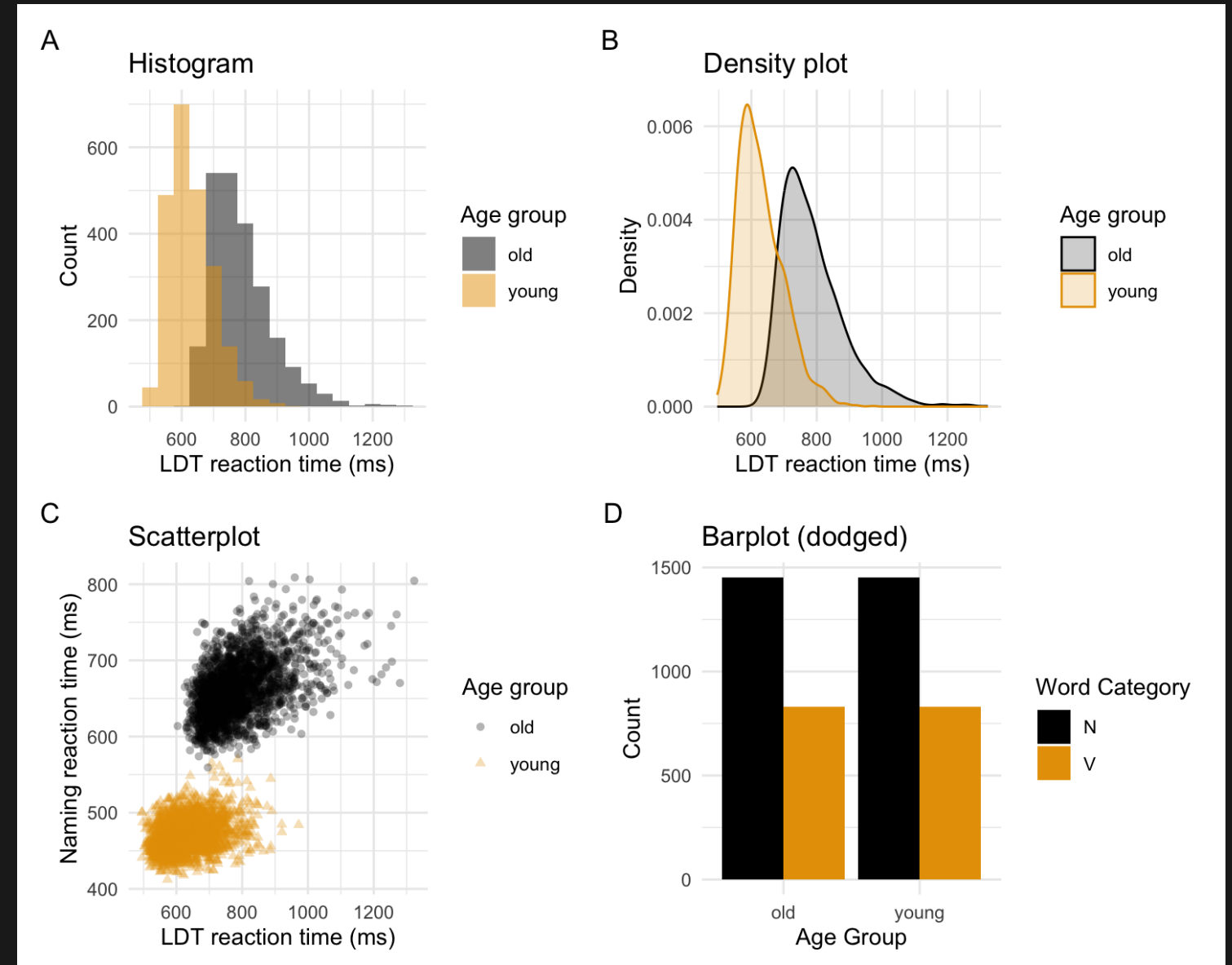
Daten

```
1 df_eng <- read_csv(  
2   here(  
3     "daten",  
4     "languageR_english.csv"  
5   )  
6 ) |>  
7 clean_names() |>  
8 rename(  
9   rt_lexdec = r_tlexdec,  
10  rt_naming = r_tnaming  
11 )
```

Wiederholung

- Betrachten Sie jede Abbildung in [Abbildung 1](#)
 - Wie viele Variablen werden in jeder Abbildung dargestellt?
 - welche *Typen* von Variablen sind es?
 - Welche zusammenfassende(n) Statistik(en) wird/werden in jedem Diagramm dargestellt?

Abbildung 1: Different plots types

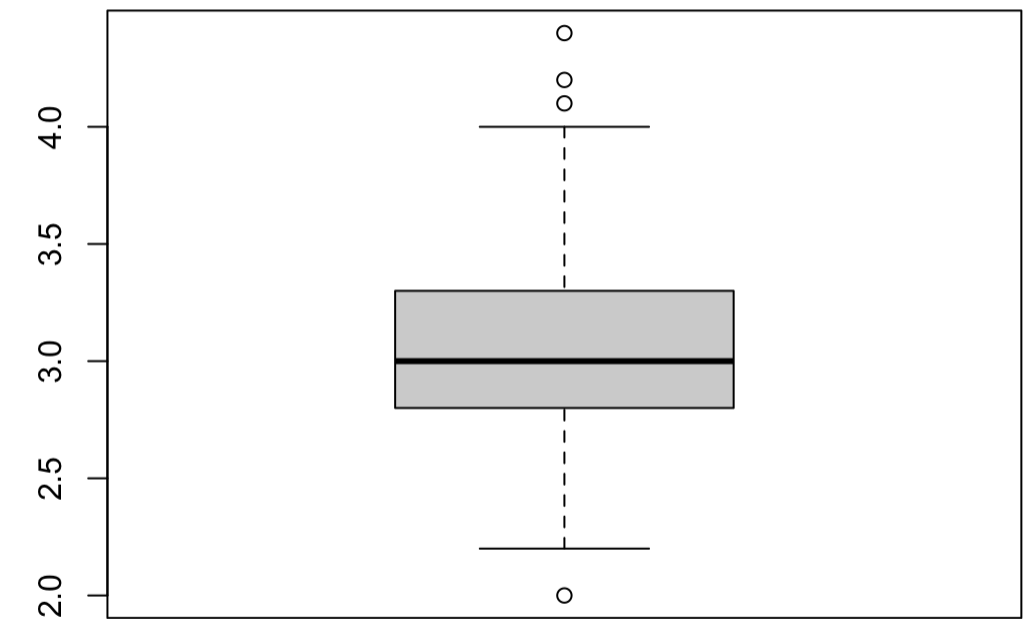


Darstellung von zusammenfassenden Statistiken

- Modus und Bereich werden in Histogrammen und Dichteplots visualisiert
- die Anzahl der Beobachtungen wird in Balkendiagrammen visualisiert

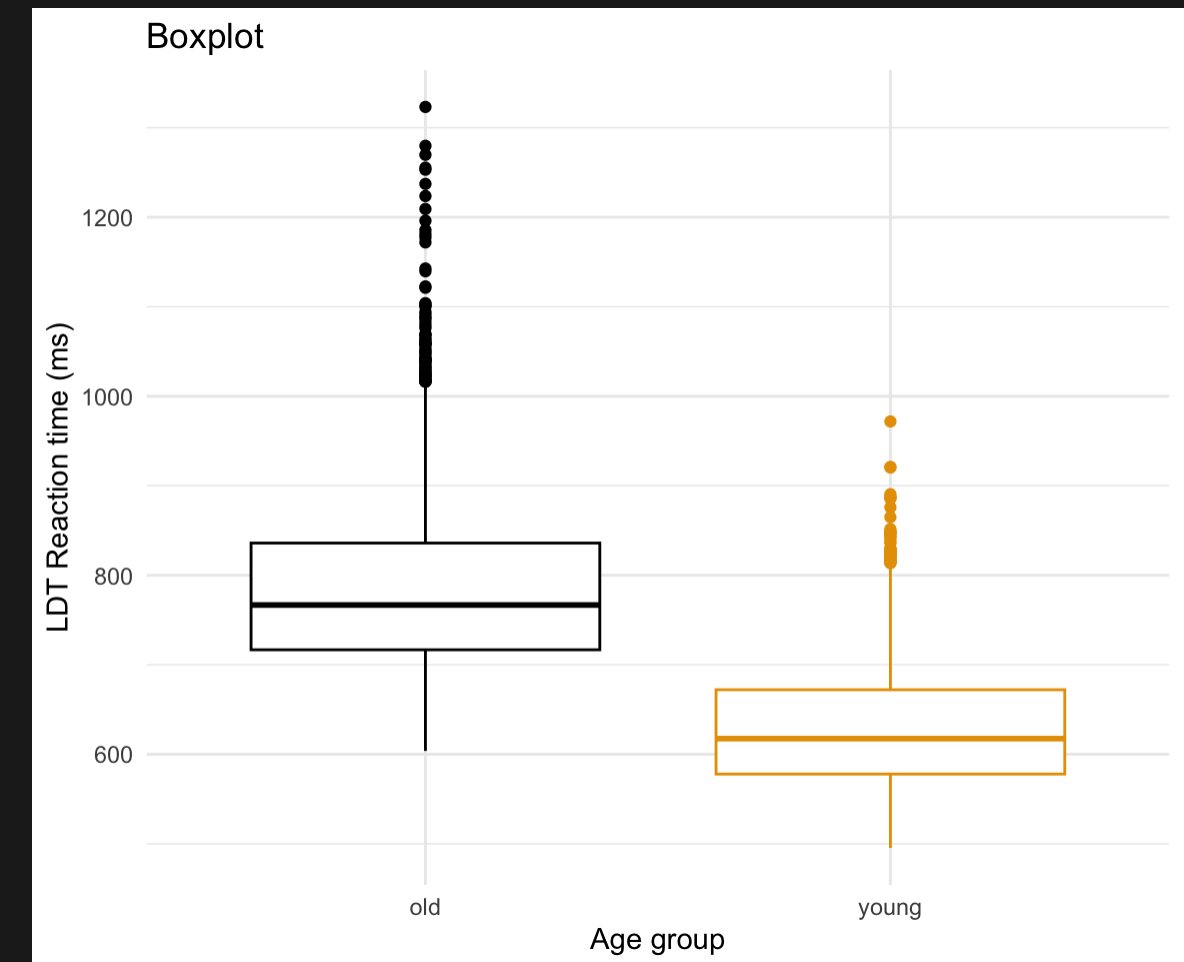
Boxplot

- auch bekannt als Box-and-Whisker-Plots, enthalten
 - eine Box
 - eine Linie in der Mitte der Box
 - Linien, die an beiden Enden der Box herausragen (die 'Whisker')
 - manchmal Punkte



- Betrachten Sie [Abbildung 2](#)
 - identifiziere jeden dieser 4 Aspekte des Plots
 - können Sie erraten, was jeder dieser Aspekte darstellen könnte und wie Sie die Darstellung interpretieren sollten?

Abbildung 2: Boxplot of `df_eng` (body mass by `age_subject`)



- Boxplots vermitteln eine Menge Informationen in einer einzigen Visualisierung
 - Die Box selbst stellt den *Interquartilsbereich* (IQR; der Bereich der Werte, der zwischen den mittleren 50% der Daten liegt) dar.
 - Die Grenzen der Box repräsentieren Q1 (1. Quartil, unter dem 25% der Daten liegen) und Q3 (3. Quartil, über dem 25% der Daten liegen)
 - die Linie in der Mitte des Boxplots stellt den *Median* dar
 - auch Q2 genannt (2. Quartil; der mittlere Wert, über/unter dem 50% der Daten liegen)
 - Die Whisker repräsentieren $1,5 * IQR$ von Q1 (unterer Whisker) oder Q3 (oberer Whisker)
 - Punkte, die außerhalb der Whisker liegen, stellen Ausreißer dar (d. h. Extremwerte, die außerhalb des IQR liegen).

- [Abbildung 3](#) zeigt die Beziehung zwischen einem Histogramm und einem Boxplot

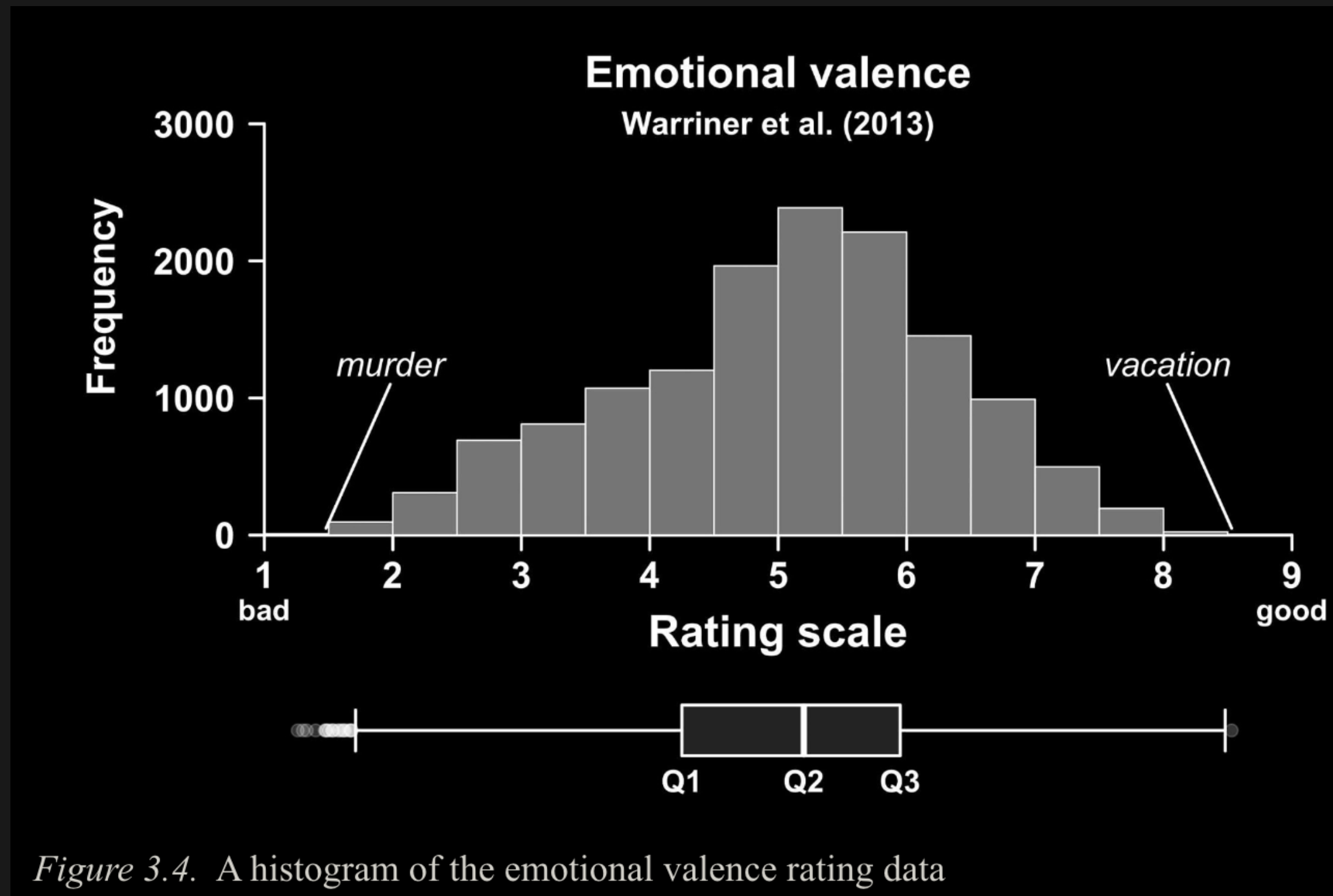


Abbildung 3: Image source: Winter ([2019](#)) (all rights reserved)

- [Abbildung 4](#) hat einen ähnlichen Vergleich, einschließlich eines Streudiagramms

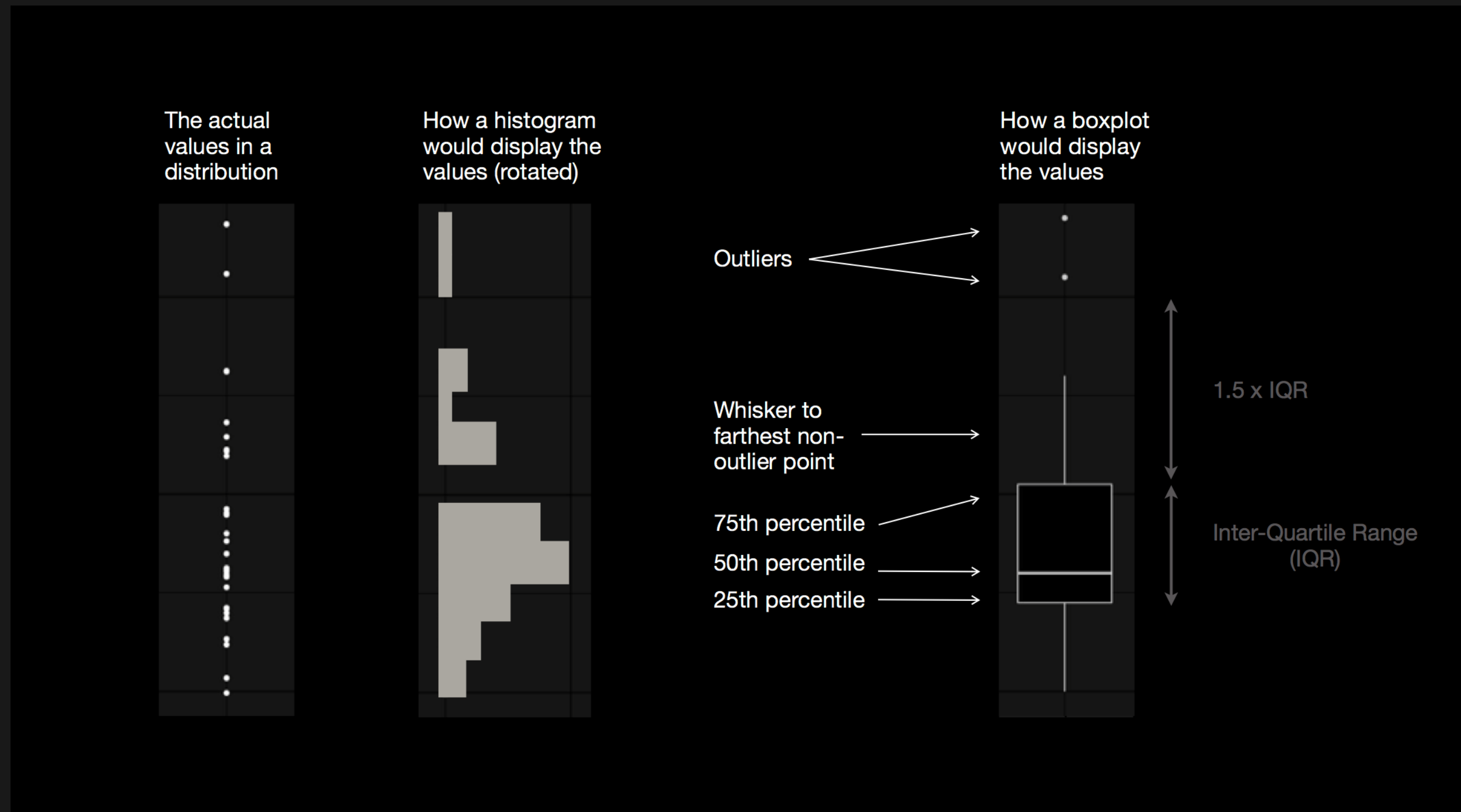


Abbildung 4: Image source: Wickham et al. (2023) (all rights reserved)

geom_boxplot()

- Die Funktion `geom_boxplot()` von `ggplot2` erzeugt Boxplots
 - sie benötigt eine numerische Variable als `x` oder `y` Achse ([Abbildung 5](#))

```
1 df_eng |>  
2   ggplot(aes(y = rt_lexdec)) +  
3   geom_boxplot()
```

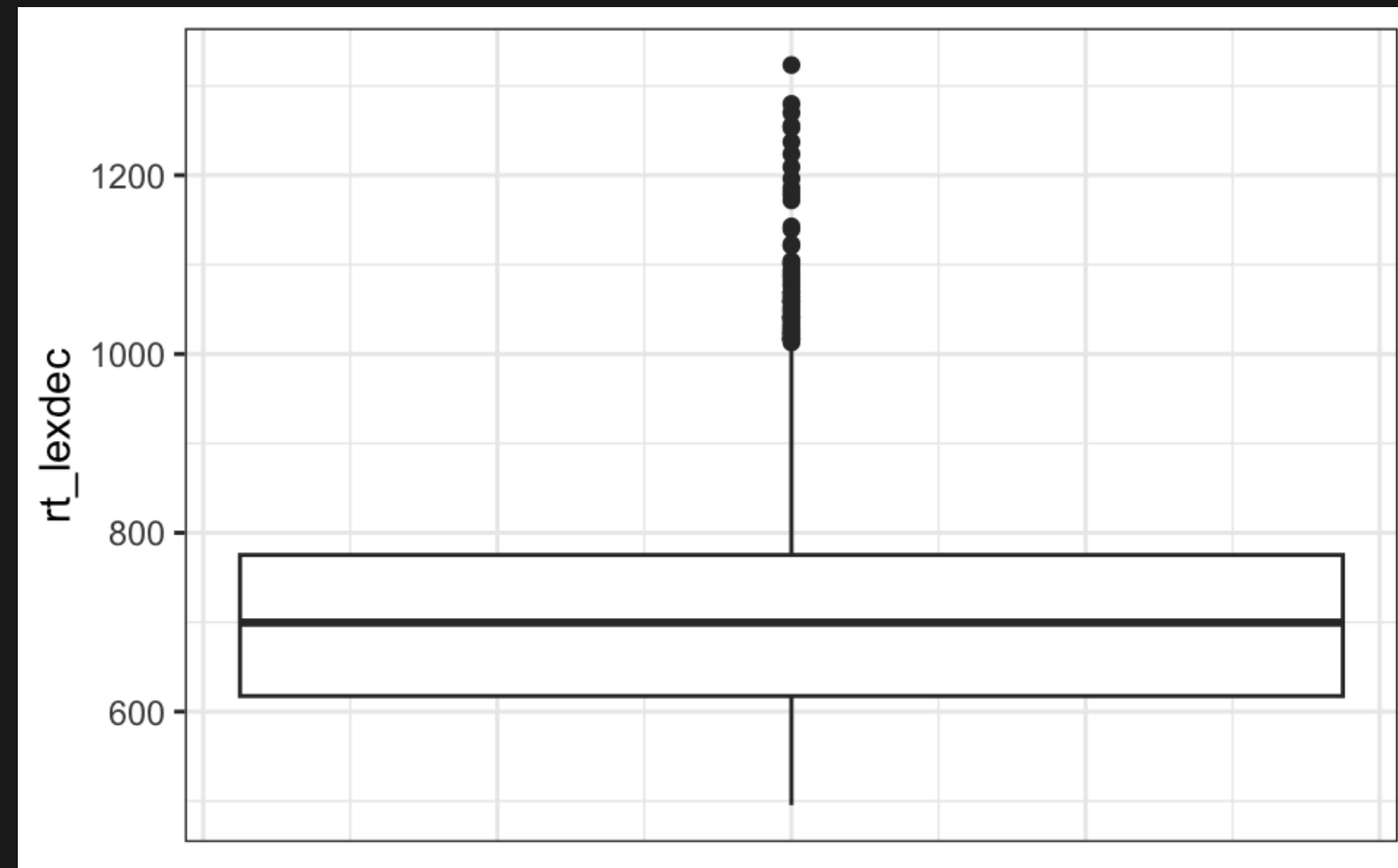


Abbildung 5: A boxplot for all observations of a continuous variable

- für Boxplots verschiedener Gruppen: eine kategoriale Variable entlang der anderen Achse (Abbildung 6)

```
1 df_eng |>  
2   ggplot(aes(x = age_subject, y = rt_lexdec)) +  
3   geom_boxplot() +  
4   theme_bw()
```

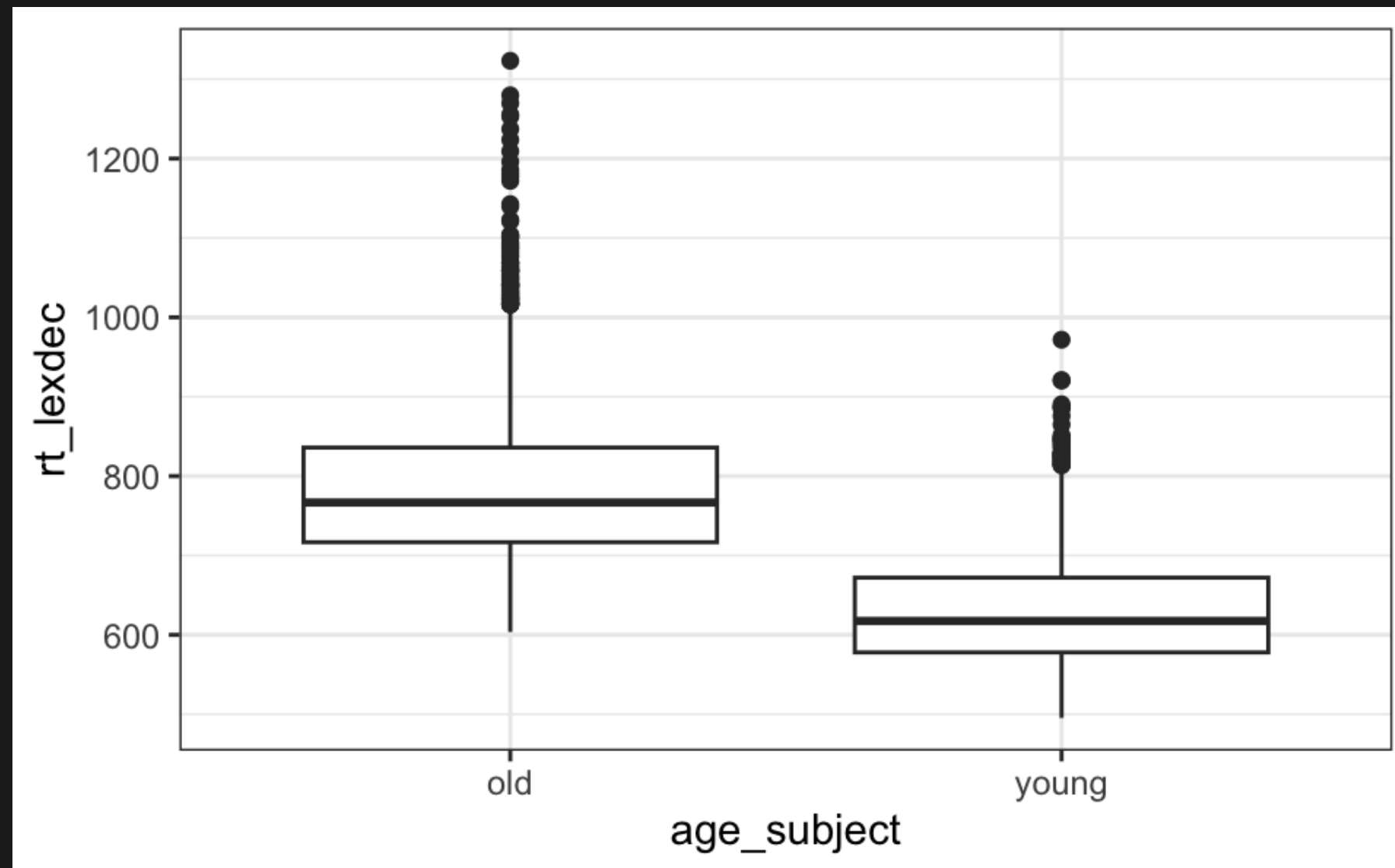


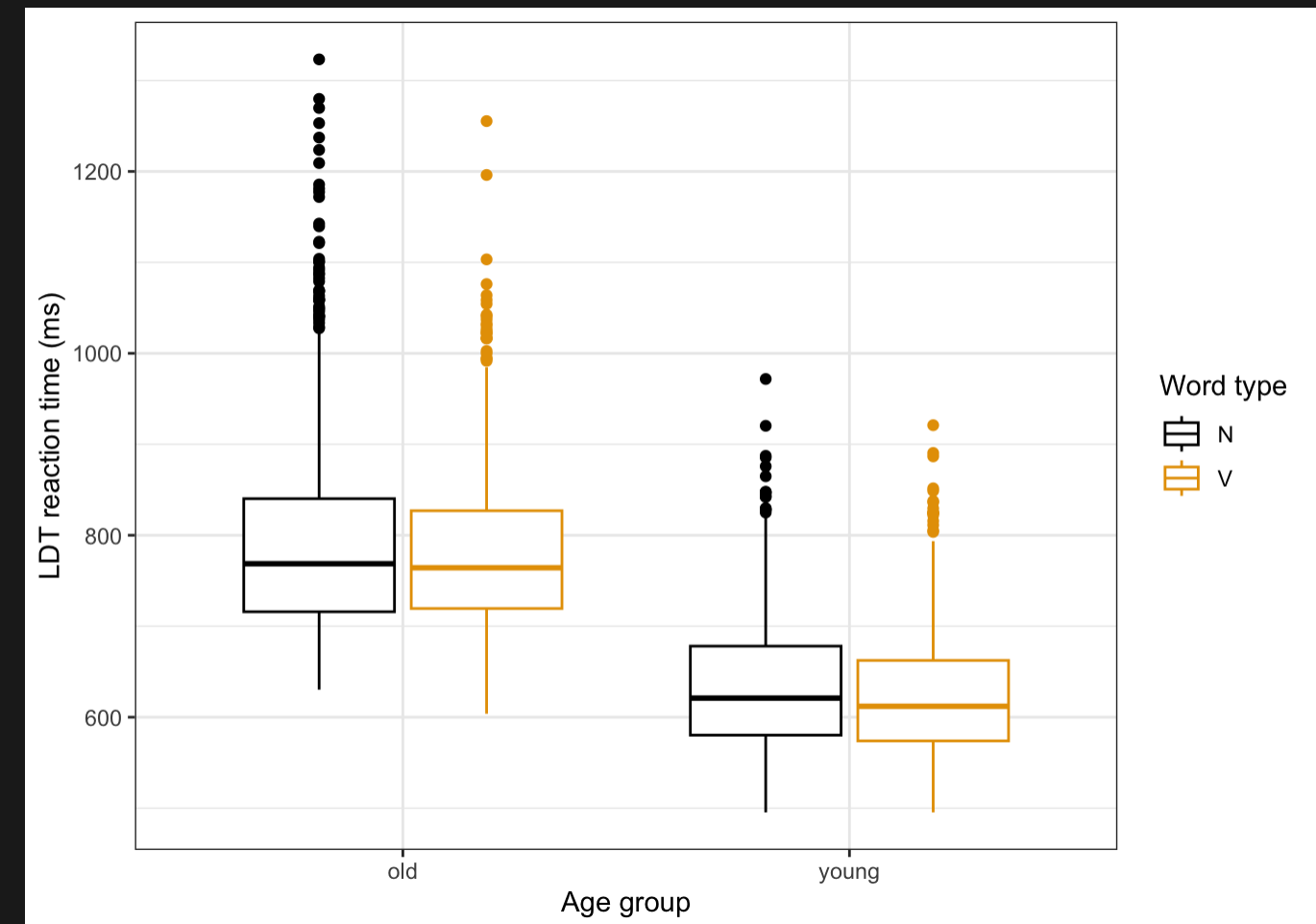
Abbildung 6: A boxplot for two groups

Gruppiertes Boxplot

- Wir können gruppierte Boxplots erstellen, um mehr Variablen zu visualisieren
 - einfach eine neue Variable mit **colour** oder **fill** ästhetisch zuordnen

```
1 df_eng |>
2   ggplot(aes(x = age_subject, y = rt_lexdec,
3             colour = word_category)) +
4   geom_boxplot() +
5   labs(
6     x = "Age group",
7     y = "LDT reaction time (ms)",
8     color = "Word type"
9   ) +
10  scale_colour_colorblind() +
11  theme_bw()
```

A grouped boxplot



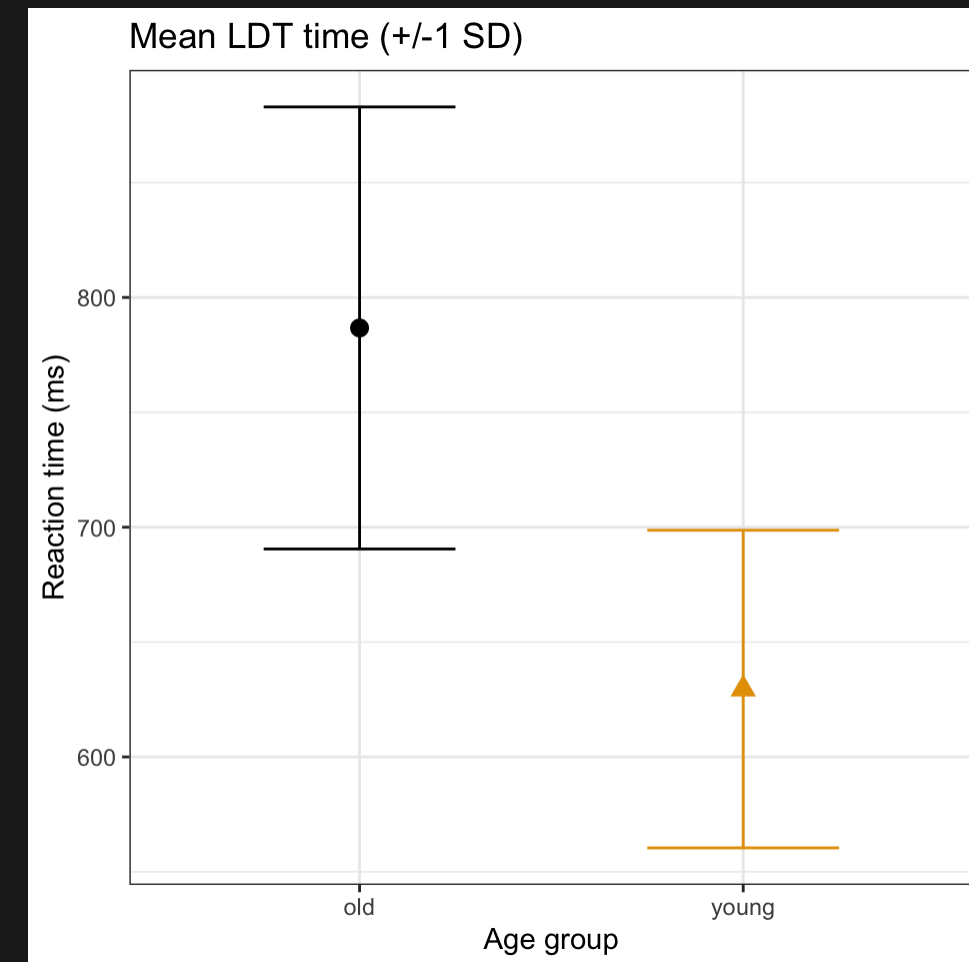
Visualisierung des Mittelwerts

- In der Regel wollen wir auch den Mittelwert mit der Standardabweichung darstellen.
 - Wie können wir das tun?

Fehlerbalkenplots

- Diese Diagramme bestehen aus 2 Teilen:
 - der Mittelwert, visualisiert mit `geom_point()`
 - ein Maß für die Streuung, visualisiert mit “`geom_errorbar()`”.
- für diesen Kurs werden wir die Standardabweichung verwenden
- [Abbildung 7](#) ist das, was wir heute erzeugen werden

Abbildung 7: Errorbar plot of `df_eng` (body mass by age_subject)



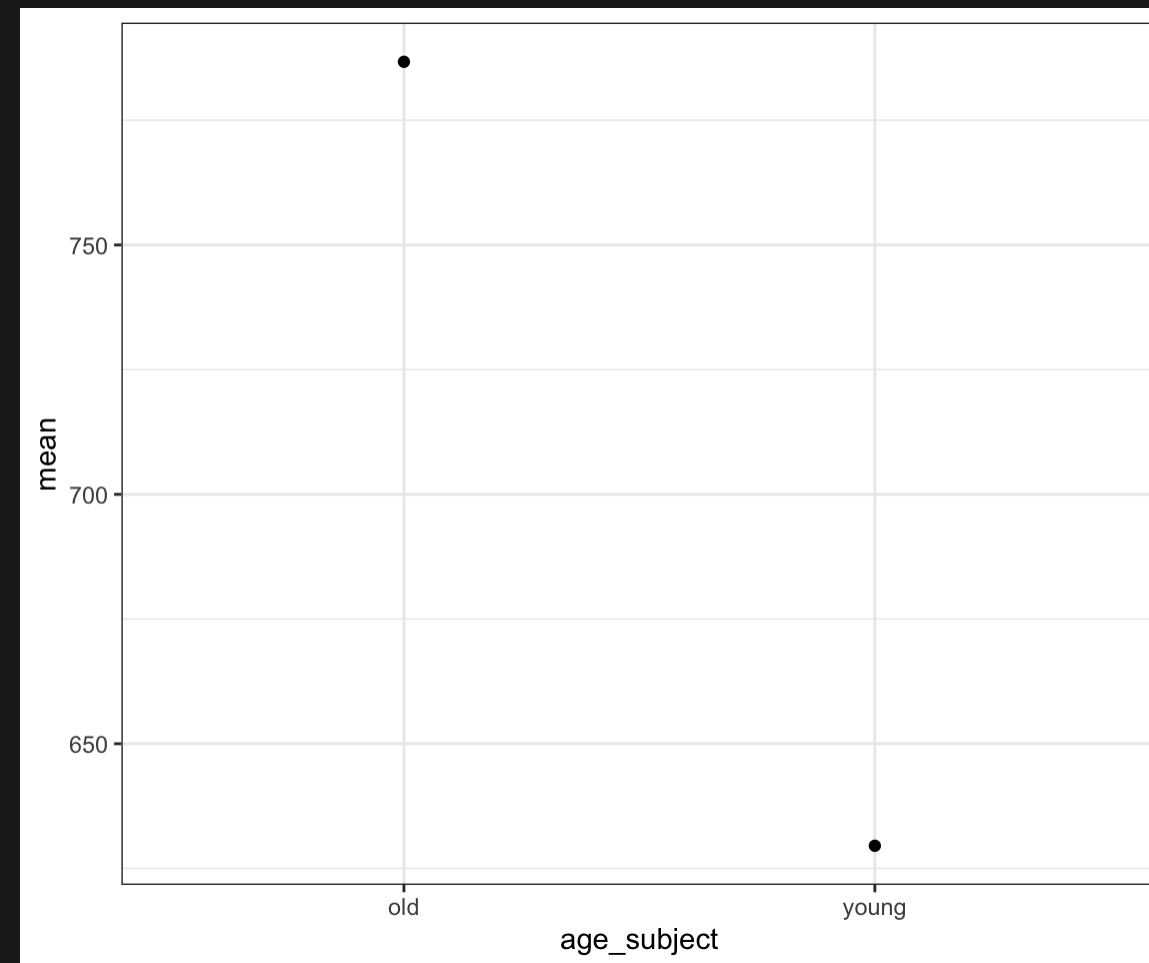
Berechnung der zusammenfassenden Statistik

- müssen wir zunächst den Mittelwert und die Standardabweichung berechnen
 - gruppiert nach den Variablen, die wir visualisieren wollen
 - Wie kann man den Mittelwert und die Standardabweichung von `rt_lexdec` nach `age_subject` berechnen?
- [Click here to see how](#)
- Diese Zusammenfassung können wir dann in `ggplot()` mit den entsprechenden ästhetischen Zuordnungen und Geomen einfügen

Plotting mean

- Zunächst werden die Mittelwerte mit `geom_point()` dargestellt.

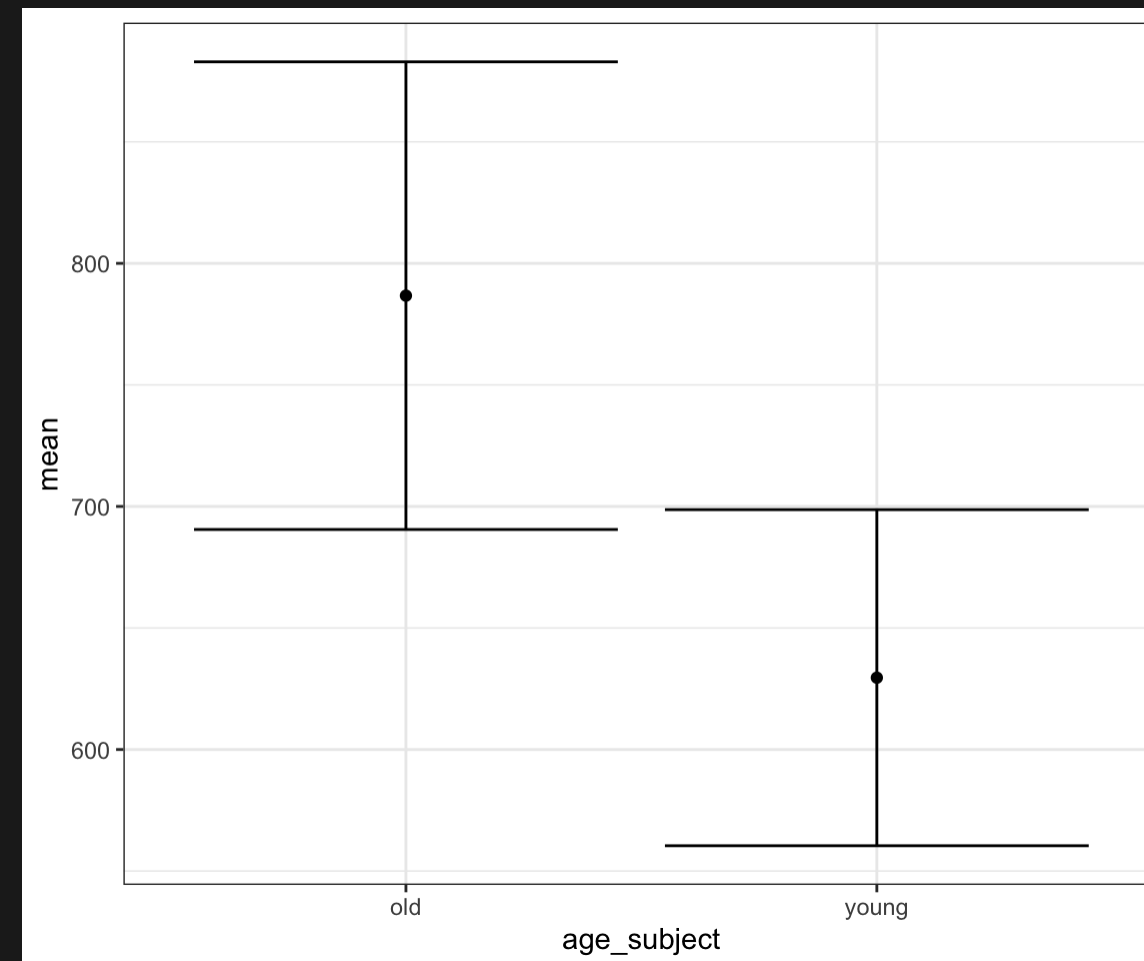
```
1 sum_eng |>  
2   ggplot() +  
3   aes(x = age_subject, y = mean) +  
4   geom_point()
```



Hinzufügen von Fehlerbalken

- Fügen wir nun unsere Fehlerbalken hinzu, die 1 Standardabweichung über und unter dem Mittelwert darstellen
- wir tun dies mit `geom_errorbar()`
 - nimmt `ymin` und `ymax` als Argumente
 - In unserem Fall sind dies `mean-sd` / `mean+sd`.

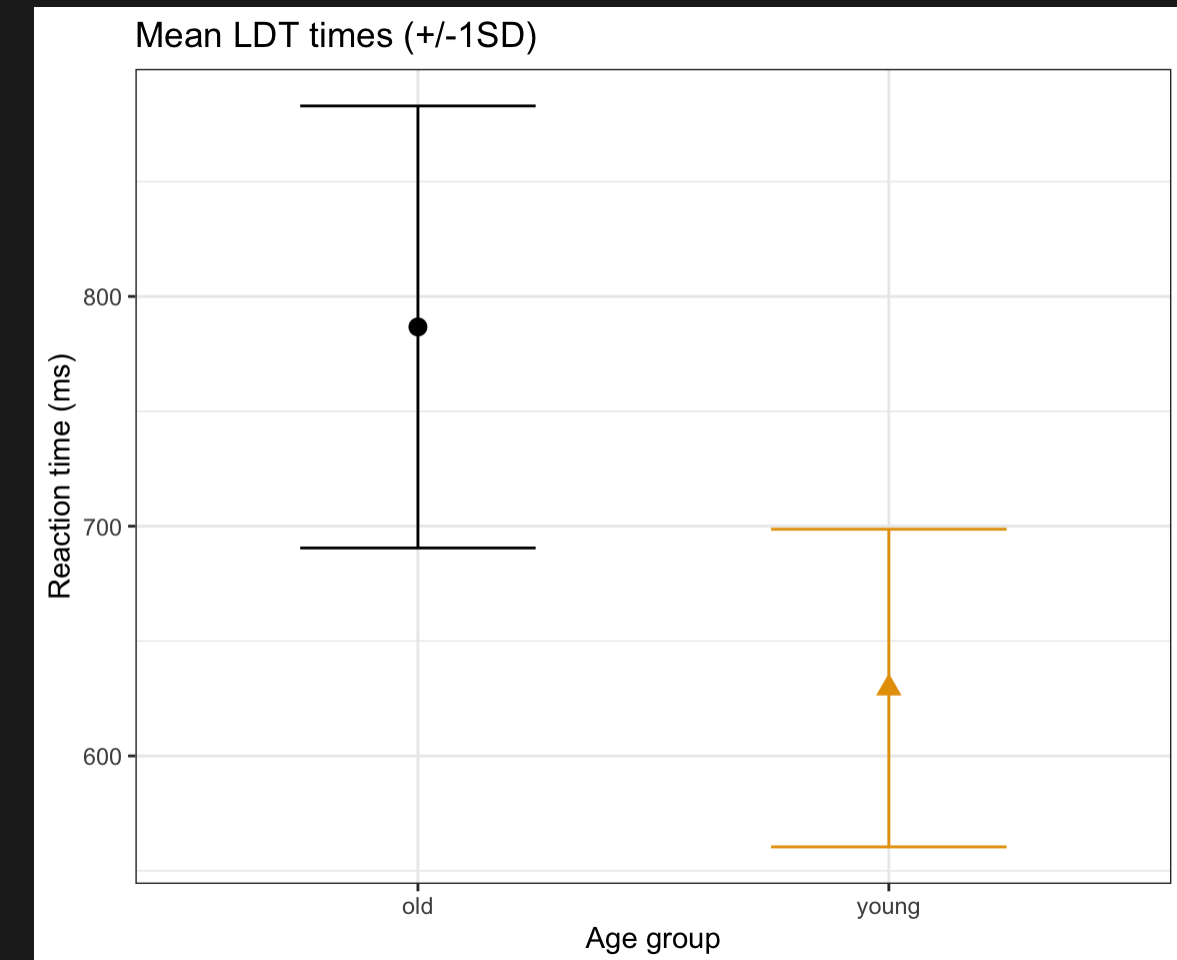
```
1 sum_eng |>
2   ggplot() +
3   aes(x = age_subject, y = mean) +
4   geom_point() +
5   geom_errorbar(aes(ymin = mean-sd,
6                     ymax = mean+sd))
```



- Wenn wir weitere Anpassungen hinzufügen, erhalten wir [Abbildung 8](#)

► Code

Abbildung 8: Customised errorbar





Barplot von Mittelwerten: Finger weg!

- Sie werden sehr oft Balkendiagramme von Mittelwerten sehen
 - aber es gibt viele Gründe, warum dies eine schlechte Idee ist!!
- Der Balkenplot hat ein schlechtes Daten-Tinten-Verhältnis, d.h. die Menge der Datentinte geteilt durch die Gesamttinte, die zur Erstellung der Grafik benötigt wird
 - Was ist, wenn es nur sehr wenige oder gar keine Beobachtungen in der Nähe von Null gibt? Wir verbrauchen eine Menge Tinte, wo es keine Beobachtungen gibt!
 - Außerdem deckt der Balken nur den Bereich ab, in dem die untere *Hälfte* der Beobachtungen liegt; ebenso viele Beobachtungen liegen über dem Mittelwert!
- Fehlerbalken allein sind keine Lösung: auch hier wird eine Menge Information verborgen
 - ein guter Grund, die Rohdatenpunkte *immer* zu visualisieren, unabhängig davon, welche zusammenfassende Darstellung Sie erstellen

Lernziele

In diesem Abschnitt haben wir gelernt, wie man...

- Boxplots erstellen und interpretieren 
- Fehlerbalkendiagramme erstellen und interpretieren 

Hausaufgabe

Boxplot mit Facette

1. Erzeugen Sie einen Plot namens `fig_boxplot`, der ein Boxplot der `df_eng` Daten ist, mit:
 - `age_subject` auf der `x`-Achse
 - `rt_naming` auf der `y`-Achse
 - `age_subject` als `colour` oder `fill` (wähle eine, es gibt keine falsche Wahl)
 - `Wort_Kategorie` in zwei Facetten mit `facet_wrap()` aufgetragen
 - die von Ihnen gewählte `theme_`-Einstellung (z.B. `theme_bw()`; für weitere Optionen siehe [hier](#))

Errorbar plot

2. Versuchen Sie, [Abbildung 9](#) zu reproduzieren. Hinweis: Sie werden die Variable `rt_naming` aus `df_eng` verwenden.

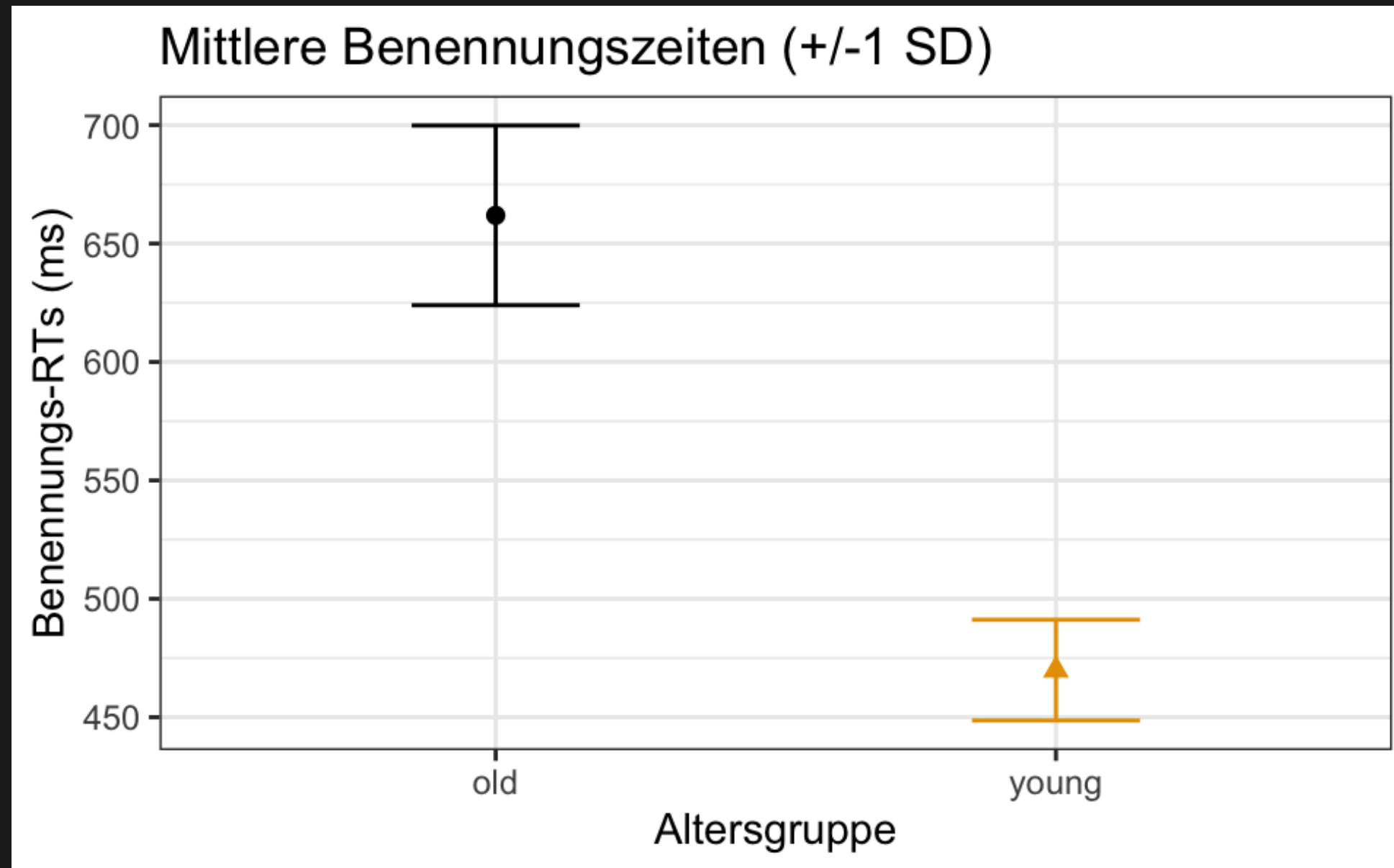


Abbildung 9: Plot to be reproduced

Patchwork

3. Verwenden Sie das Paket `patchwork`, um Ihren Boxplot und Ihre Fehlerbalkenplots nebeneinander darzustellen. Es sollte ungefähr so aussehen wie [Abbildung 10](#). Hinweis: Wenn Sie die “tag-level” (“A” und “B”) zu den Plots hinzufügen möchten, müssen Sie `+ plot_annotation(tag_level = "A")` aus `patchwork` hinzufügen.

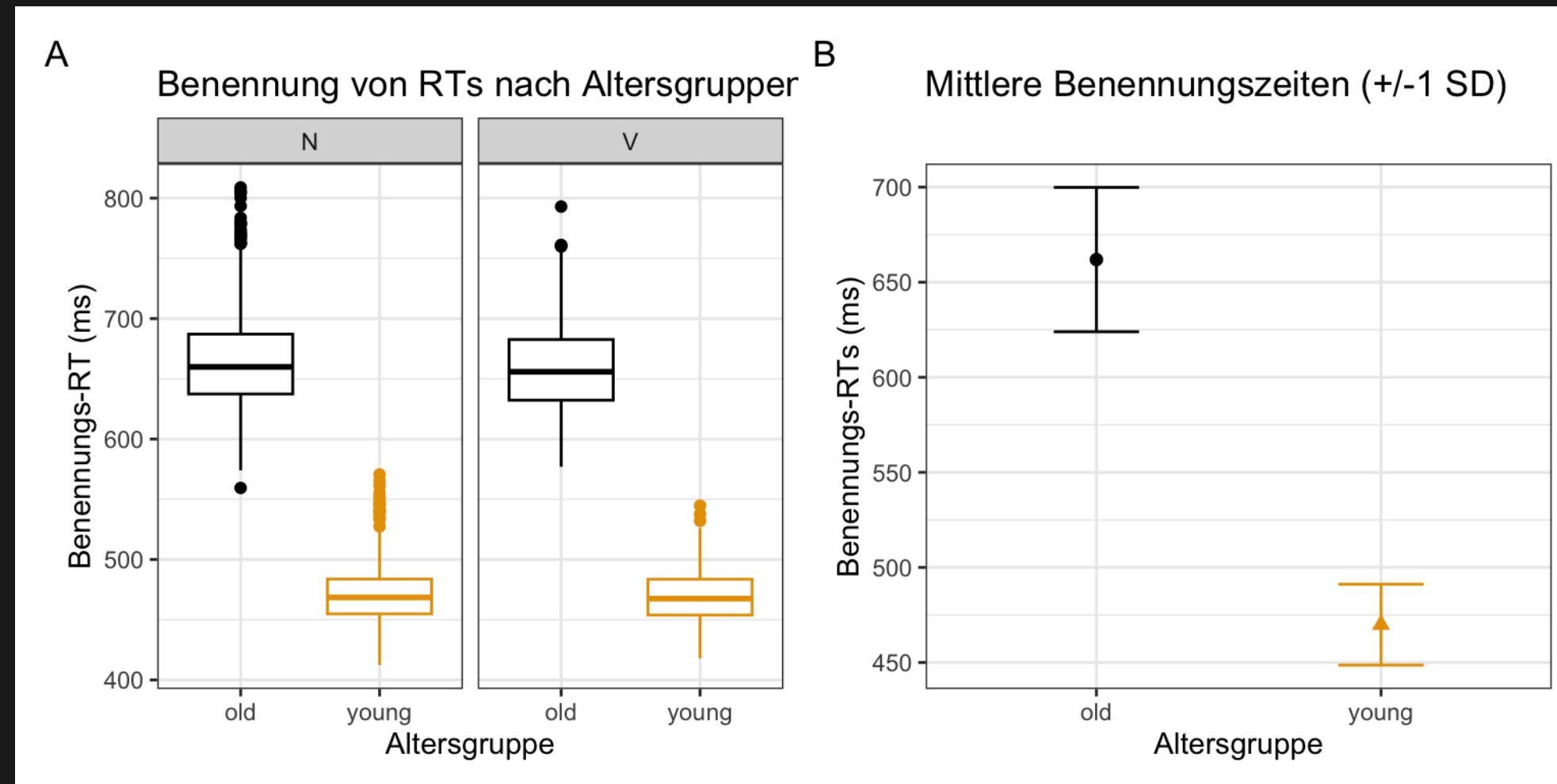


Abbildung 10: Combined plots with `patchwork`

Session Info

Hergestellt mit R version 4.3.0 (2023-04-21) (Already Tomorrow) und RStudioversion 2023.9.0.463 (Desert Sunflower).

```
1 print(sessionInfo(), locale = F)
```

R version 4.3.0 (2023-04-21)

Platform: aarch64-apple-darwin20 (64-bit)

Running under: macOS Ventura 13.2.1

Matrix products: default

BLAS: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib

LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib; LAPACK version 3.11.0

attached base packages:

[1] stats graphics grDevices utils datasets methods base

other attached packages:

```
[1] magick_2.7.4      patchwork_1.1.3  ggthemes_4.2.4   janitor_2.2.0
[5] here_1.0.1        lubridate_1.9.2  forcats_1.0.0    stringr_1.5.0
[9] dplyr_1.1.3       purrr_1.0.2      readr_2.1.4      tidyr_1.3.0
[13] ...
```

Literaturverzeichnis

- Nordmann, E., McAleer, P., Toivo, W., Paterson, H., & DeBruine, L. M. (2022). Data Visualization Using R for Researchers Who Do Not Use R. *Advances in Methods and Practices in Psychological Science*, 5(2), 251524592210746. <https://doi.org/10.1177/25152459221074654>
- Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2023). *R for Data Science* (2. Aufl.).
- Winter, B. (2019). Statistics for Linguists: An Introduction Using R. In *Statistics for Linguists: An Introduction Using R*. Routledge. <https://doi.org/10.4324/9781315165547>

