

Datenvisualisierung 3

Darstellung der zusammenfassenden Statistik

Daniela Palleschi

Humboldt-Universität zu Berlin

Lernziele

Heute werden wir lernen...

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

Ressourcen

- Kurswebsite (Datavisualisierung 3)
- 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)
```

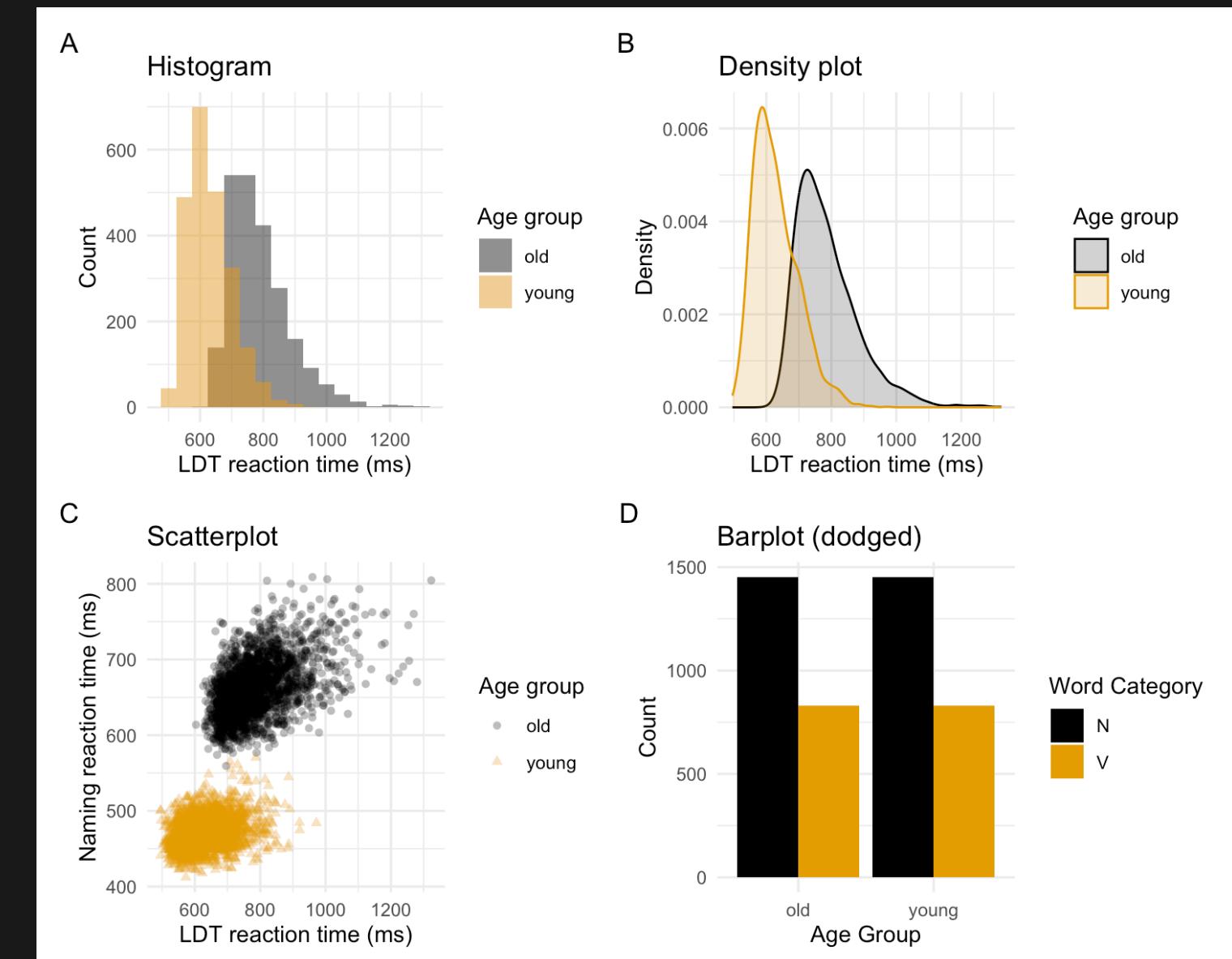
Daten

```
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

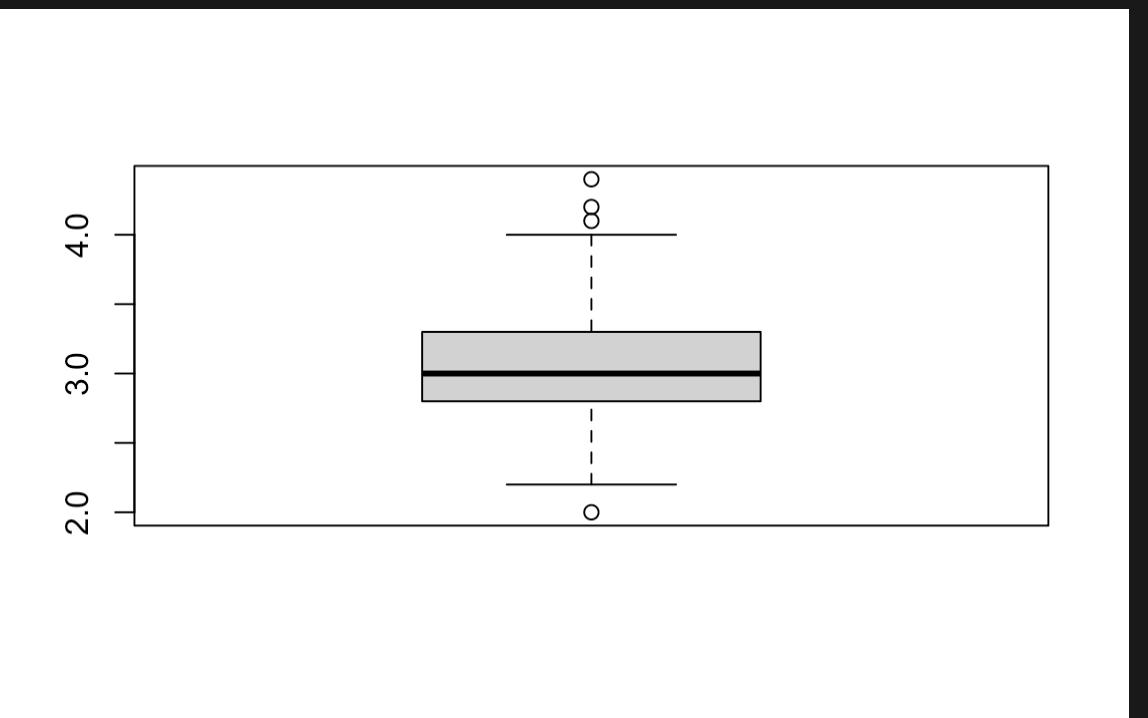


Darstellung von zusammenfassenden Statistiken

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

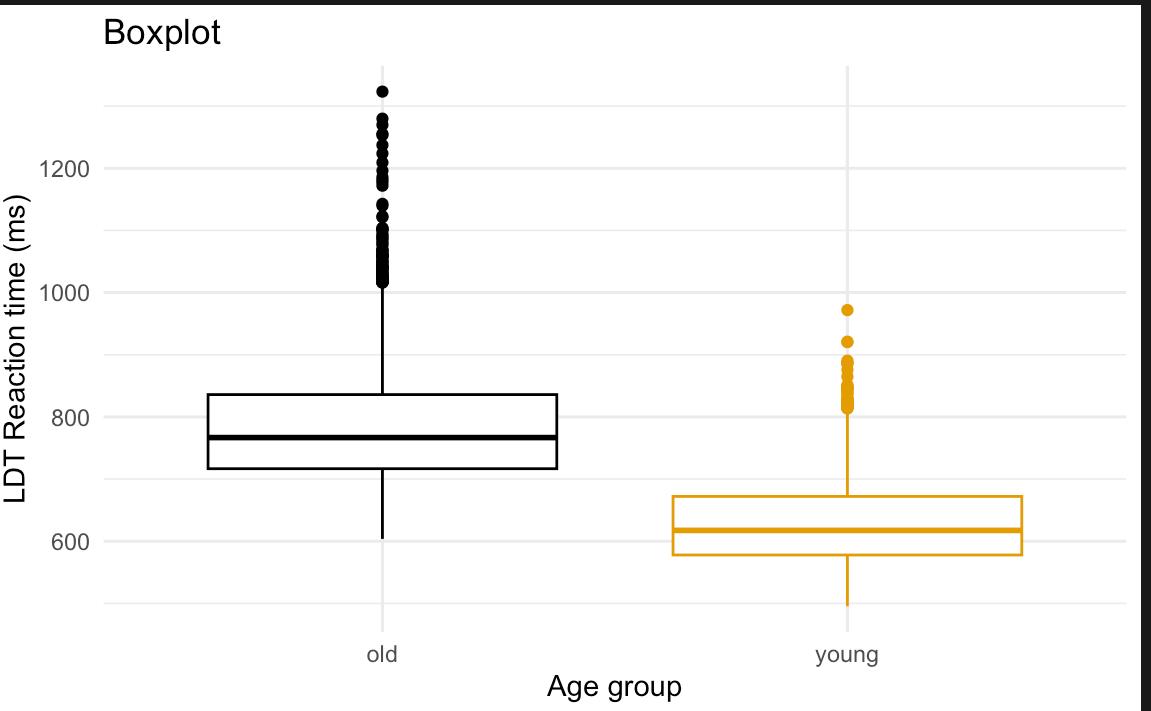
Boxplot

- 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 $1,5 * \text{IQR}$ 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

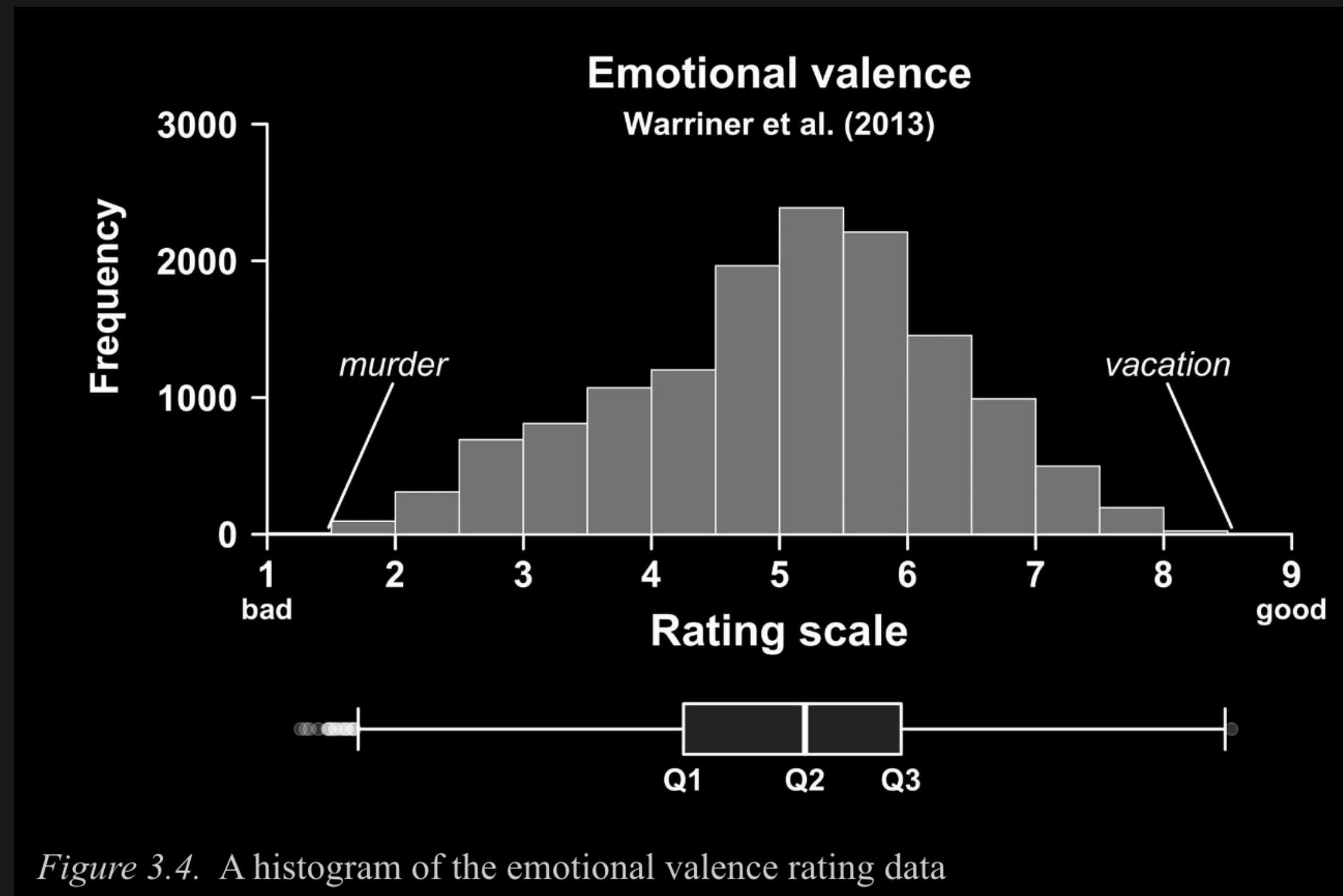


Abbildung 3: Image source: Winter (2019) (all rights reserved)

- Abbildung 4 hat einen ähnlichen Vergleich, einschließlich eines Streudiagramms

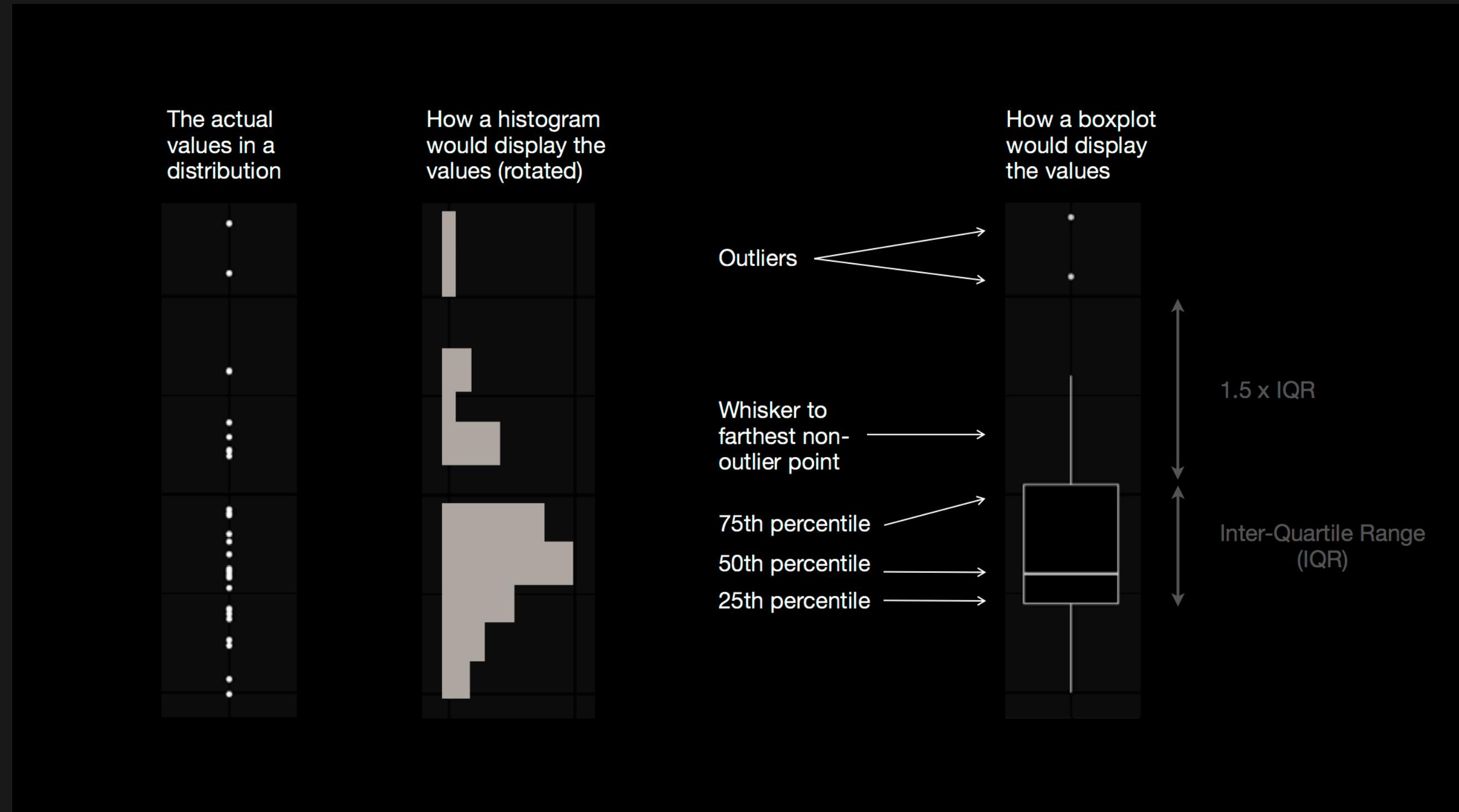


Abbildung 4: Image source: Wickham et al. (2023) (all rights reserved)

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()
```

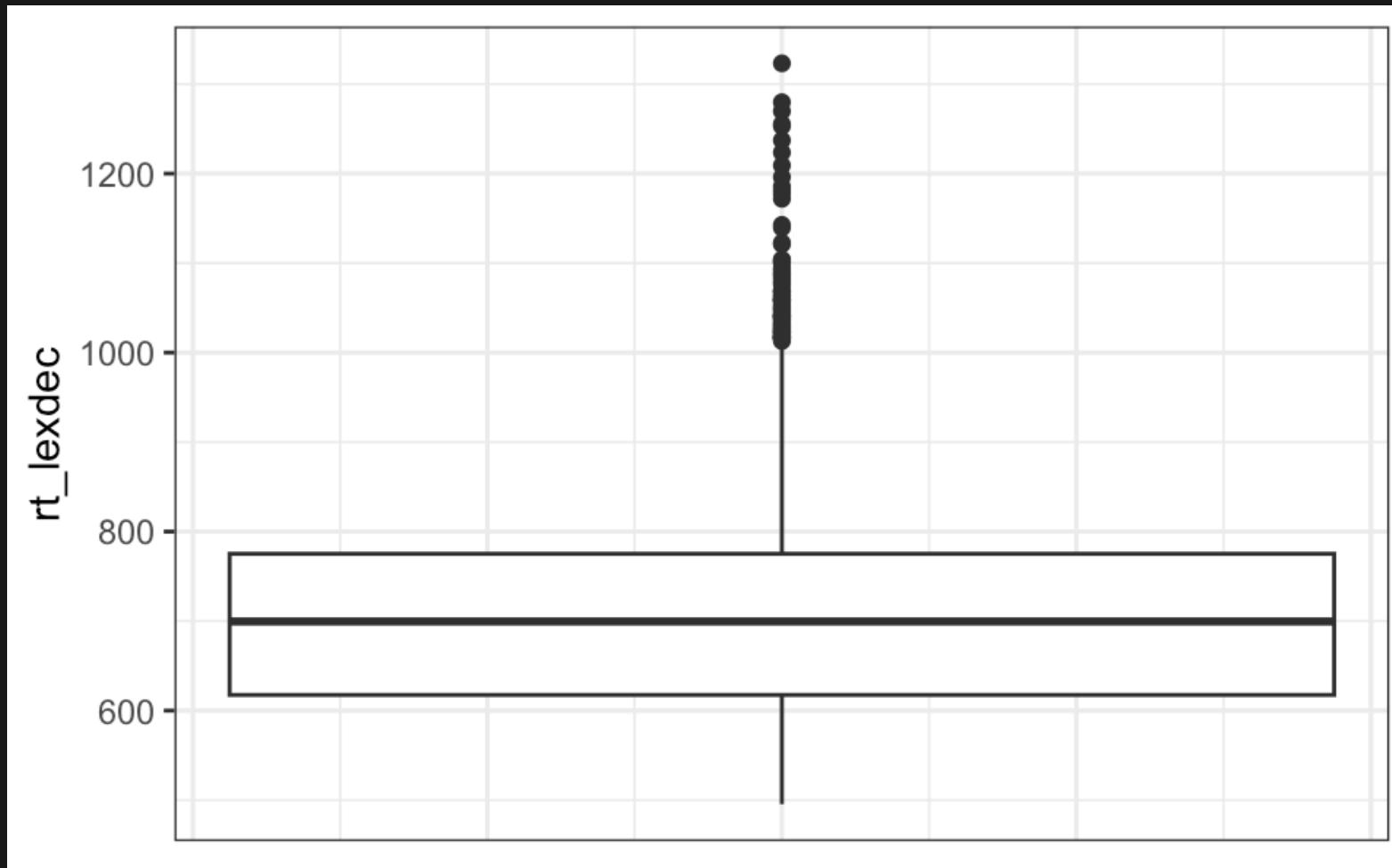


Abbildung 5: A boxplot for all observations of a continuous variable

- 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()
```

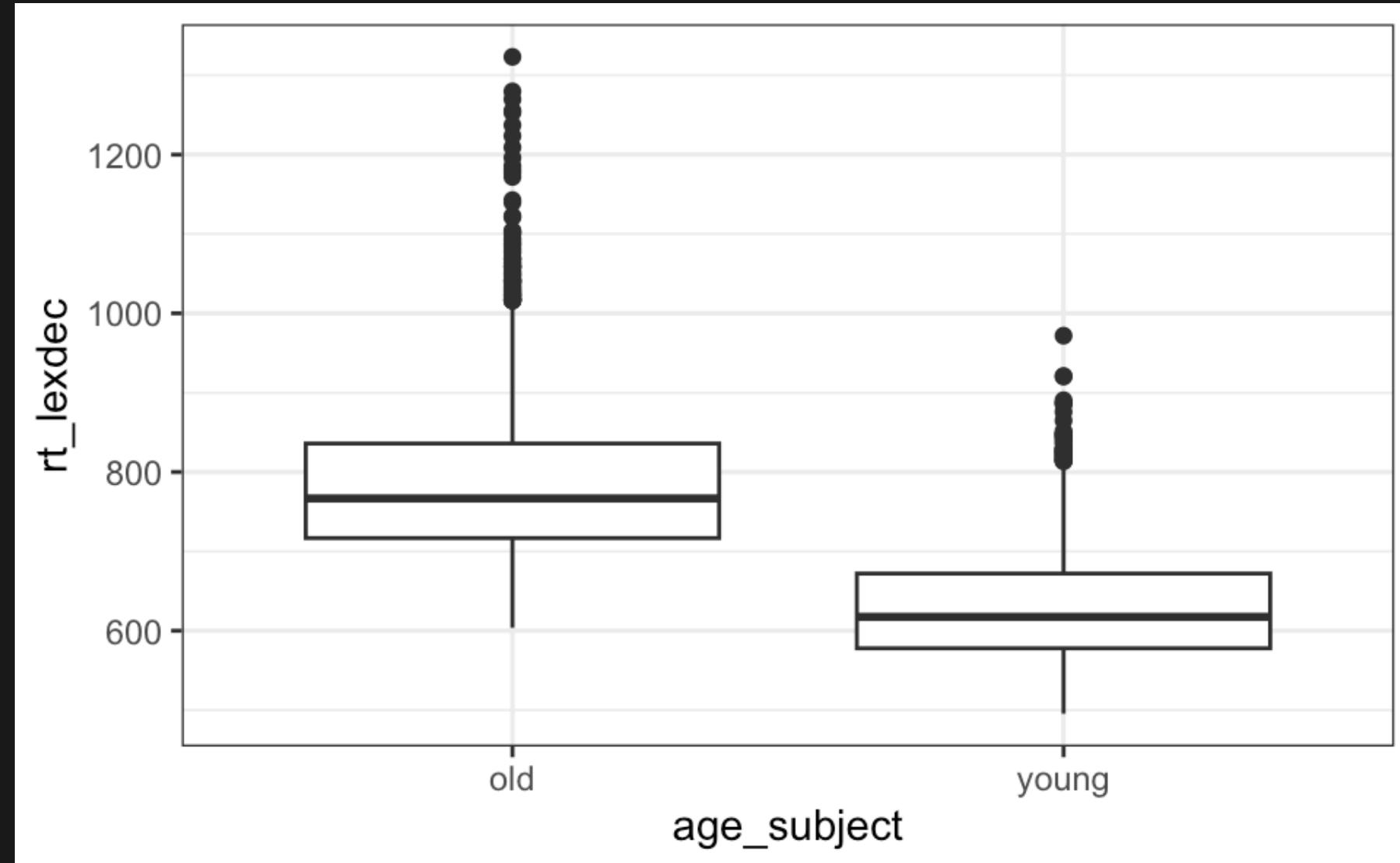


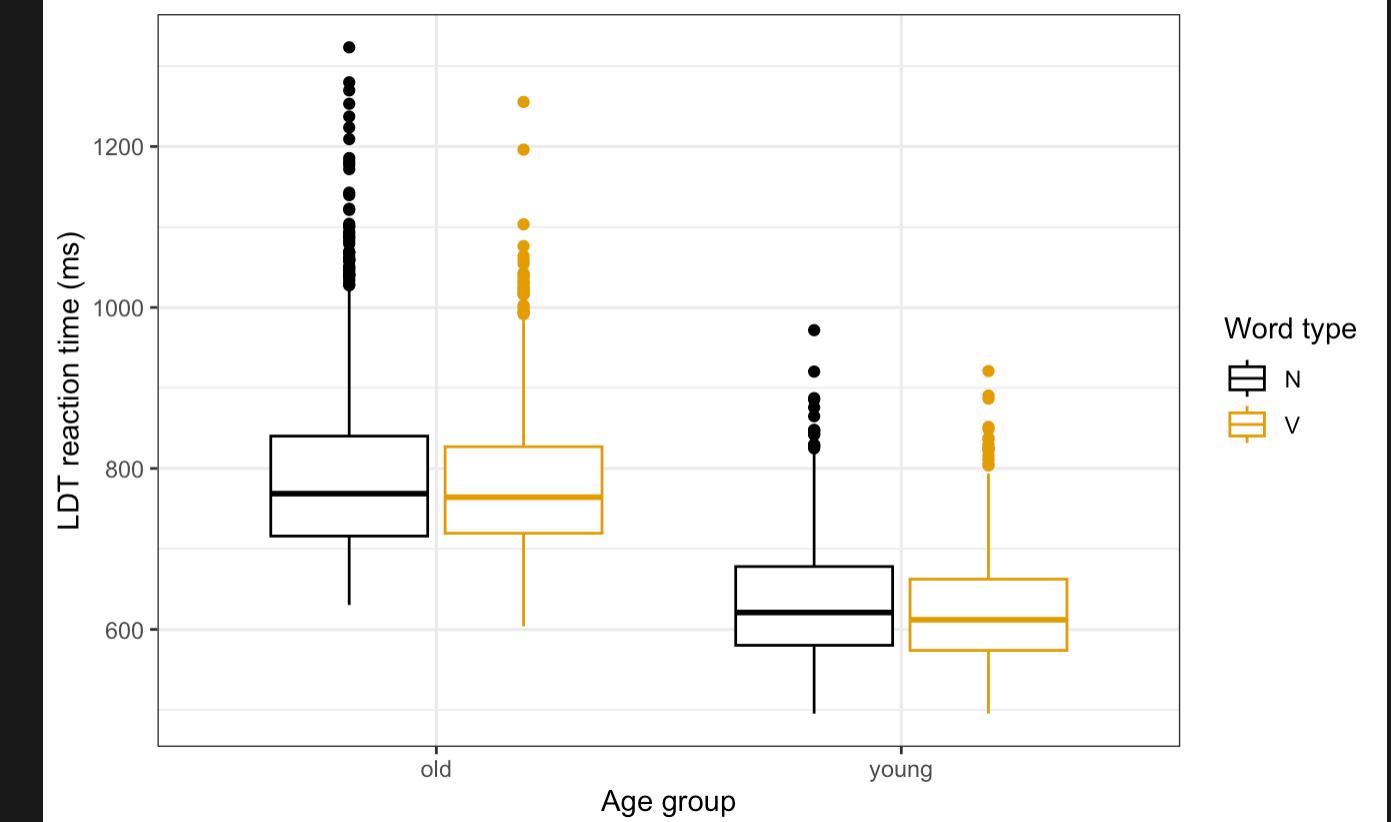
Abbildung 6: A boxplot for two groups

Gruppierter 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



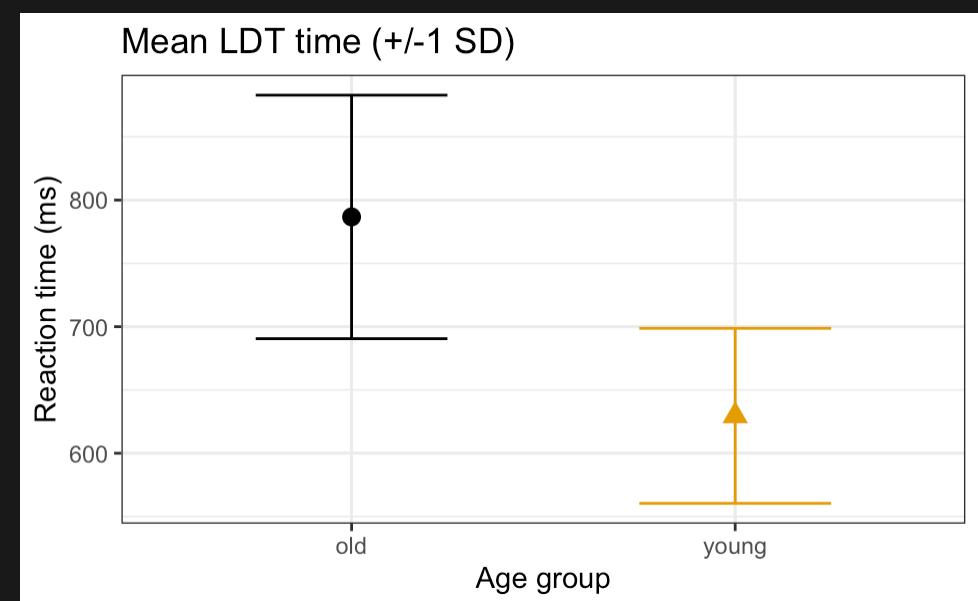
Visualisierung des Mittelwerts

- 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)



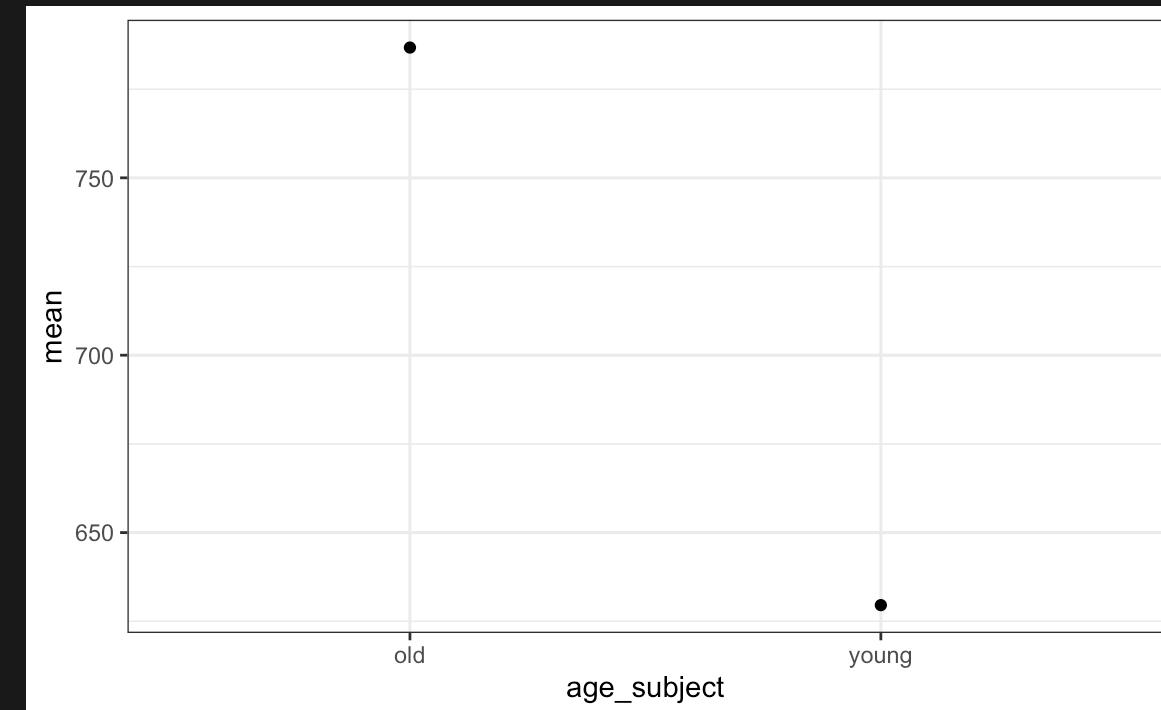
Berechnung der zusammenfassenden Statistik

- 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.

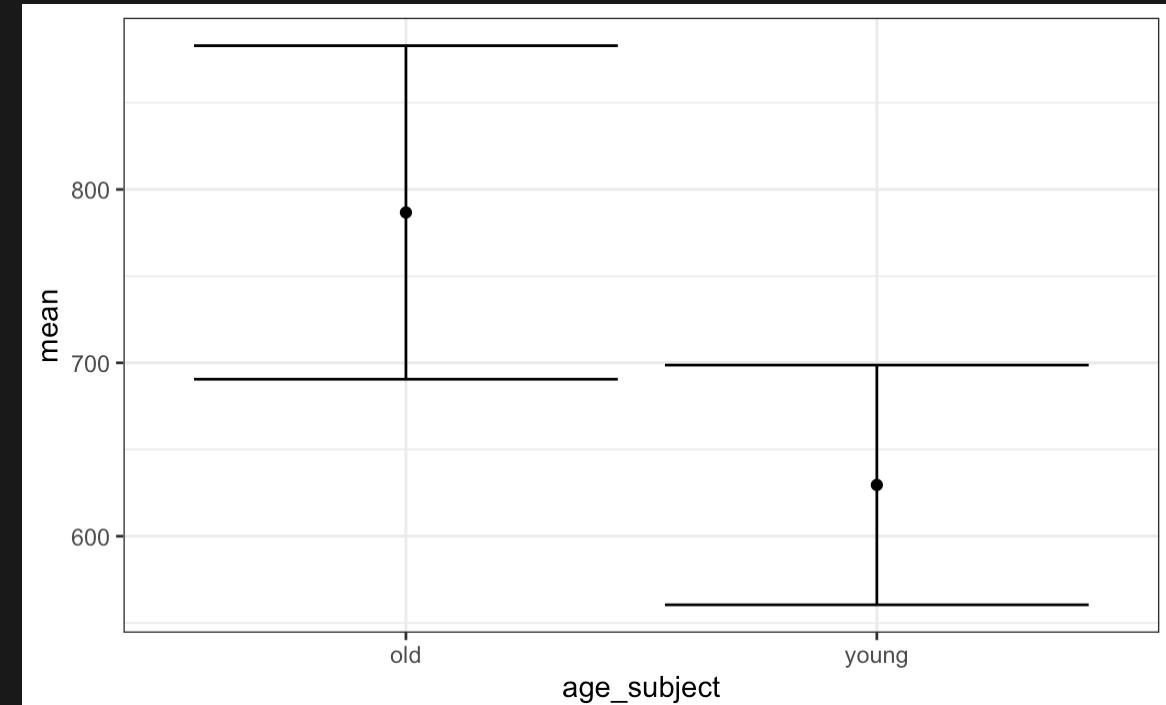
```
1 sum_eng |>
2   ggplot() +
3   aes(x = age_subject, y = mean) +
4   geom_point()
```



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`.

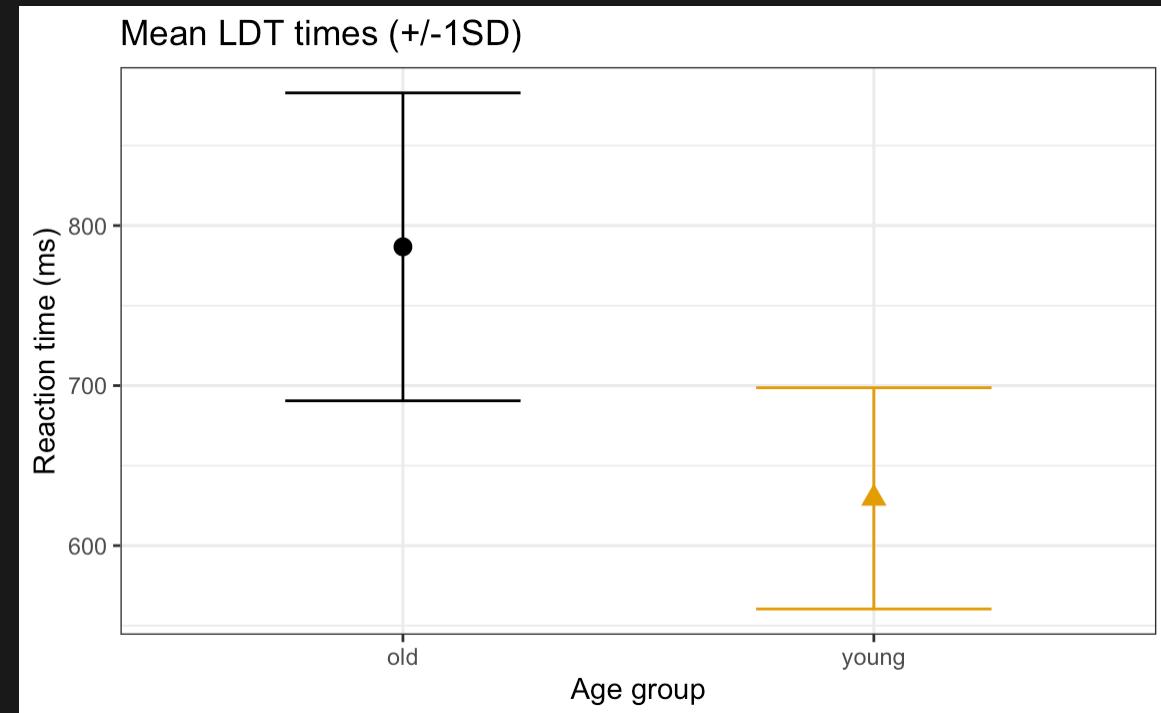
```
1 sum_eng |>
2   ggplot() +
3   aes(x = age_subject, y = mean) +
4   geom_point() +
5   geom_errorbar(aes(ymin = mean-sd,
6                      ymax = mean+sd))
```



- 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!

- Wie groß ist die Bandbreite der beobachteten Werte?

```
1 range(df_eng$rt_lexdec)  
[1] 495.38 1323.20
```

- beachten Sie, dass der tatsächliche Bereich der Datenpunkte und das Balkendiagramm viel “Tinte” für datenfreie (d. h. unbeobachtete) Reaktionszeitwerte verwenden

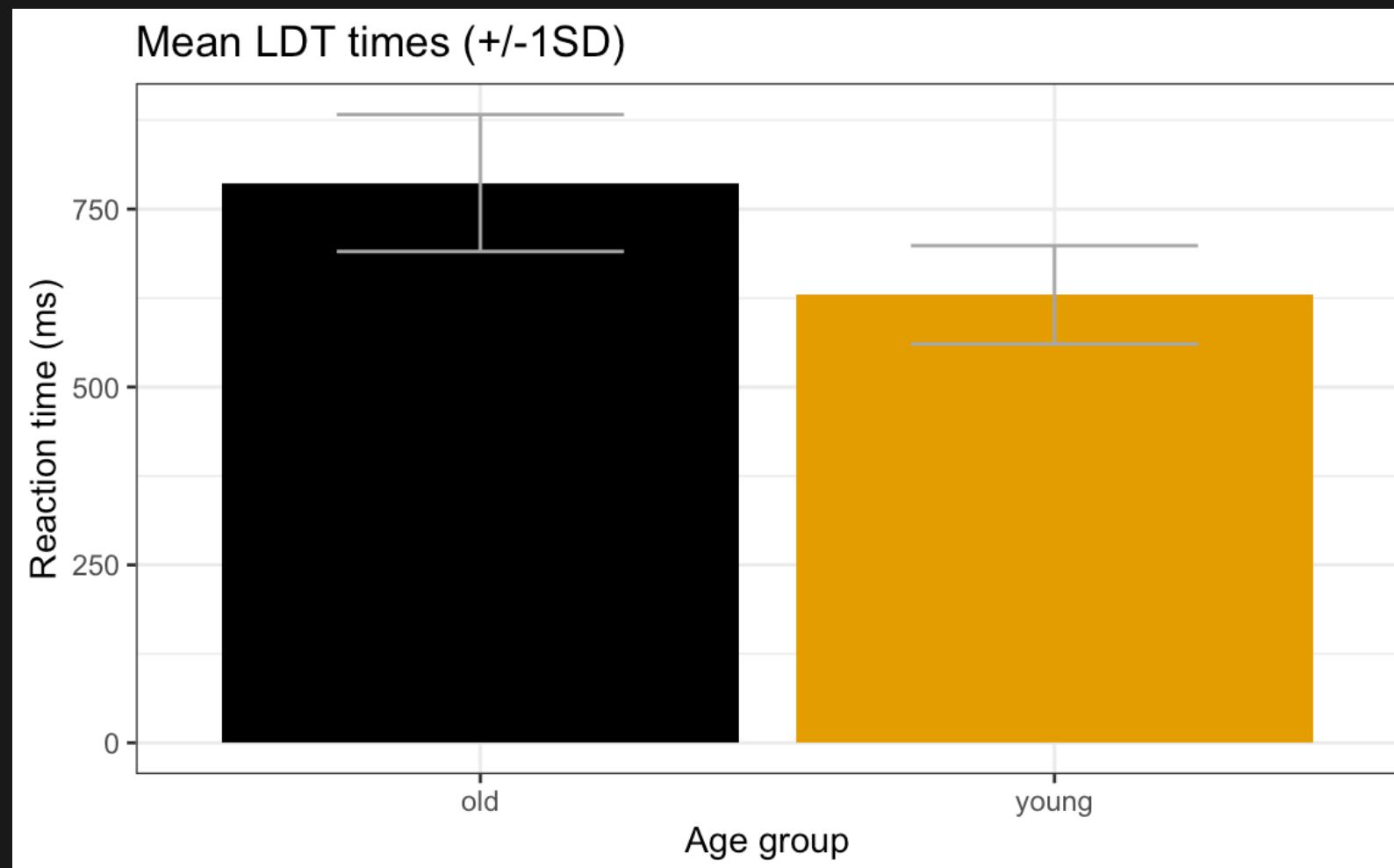
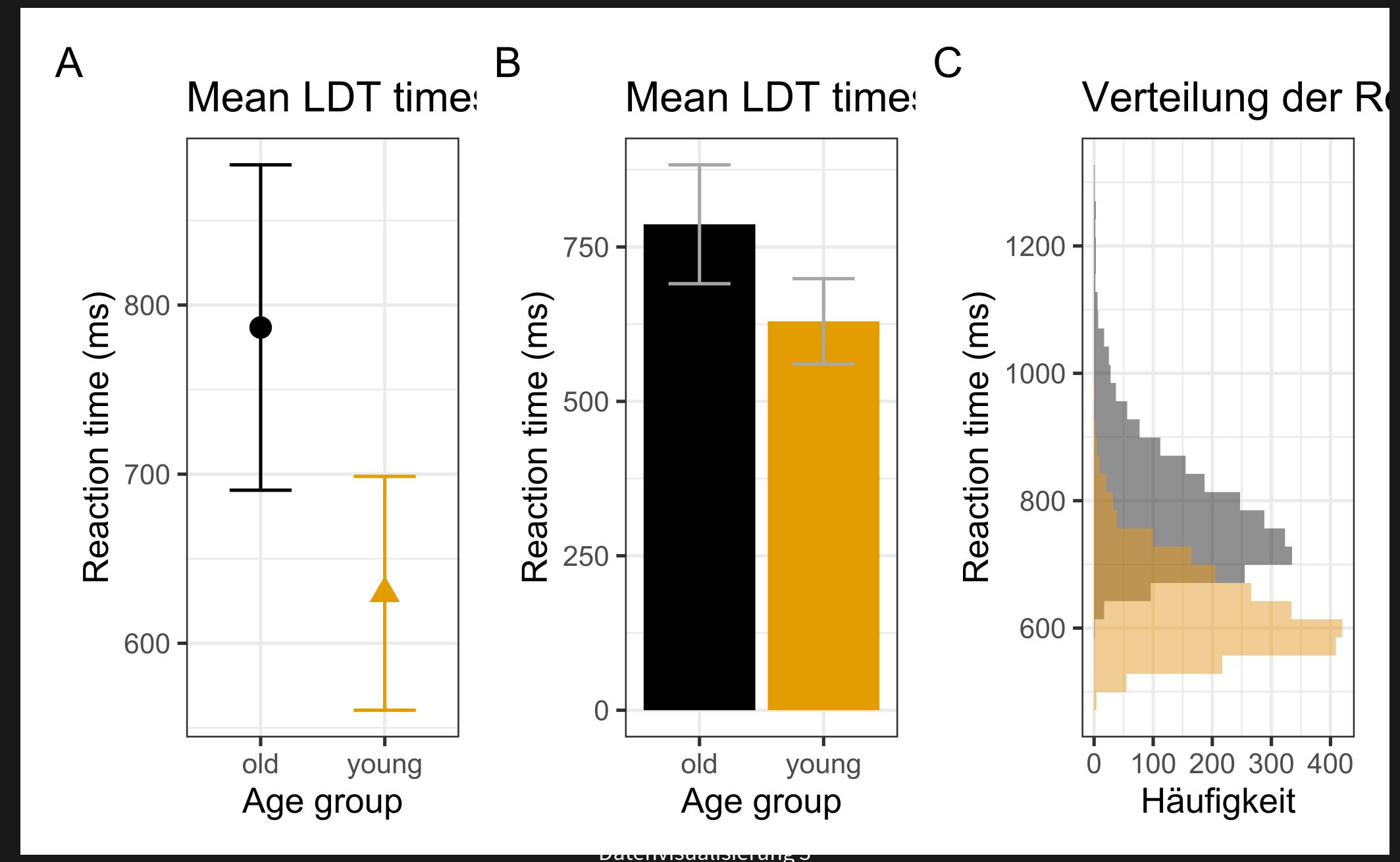


Abbildung 9: Balkendiagramm des Mittelwerts mit +/- 1 Standardabweichung (ich empfehle, von solchen Diagrammen abzusehen!)

Fehlerbalkendiagramm vs. Balkendiagramm für Mittelwerte

- Abbildung 10 A und B stellen dieselbe Information dar



Gleiche Grenzen auf der y-Achse

- Abbildung 11 zeigt die gleichen Daten, aber mit dem gleichen y-Achsenbereich

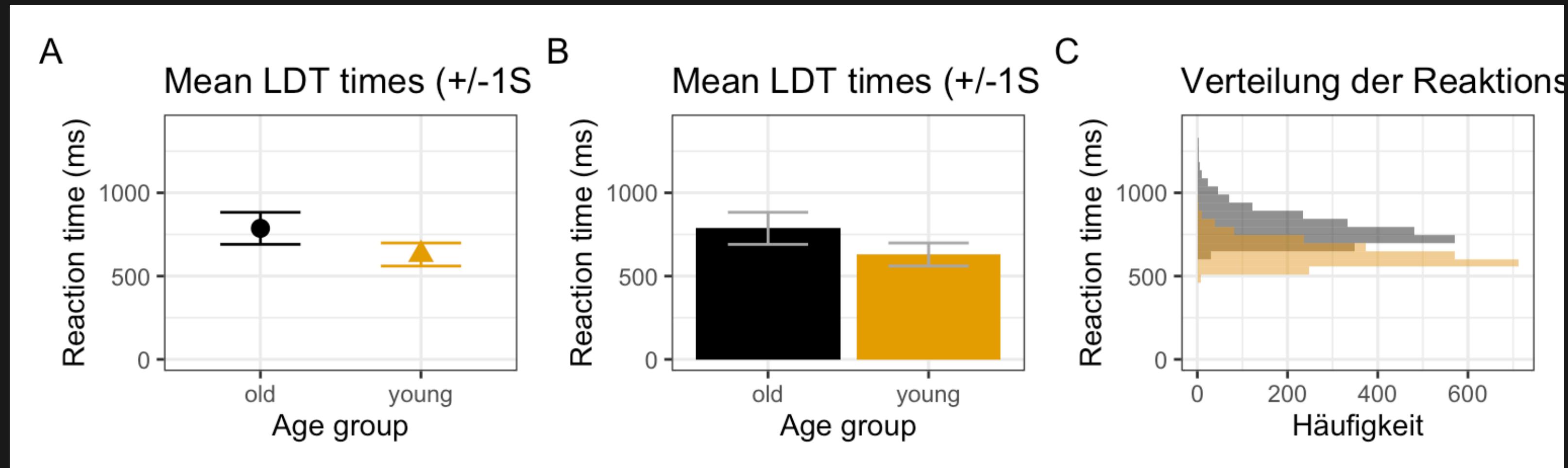


Abbildung 11: Fehlerbargraphik und Balkenplots, die Mittelwerte (+/-1SD) darstellen, sowie ein Histogramm derselben Daten mit demselben y-Achsenbereich

Zusammenfassende Statistiken und Verteilung

- 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

Anhang 8: Datenvisualisierung 3 auf der Website des Kurses.

Session Info

Hergestellt mit R version 4.4.0 (2024-04-24) (Puppy Cup) und RStudioversion 2023.9.0.463 (Desert Sunflower).

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

R version 4.4.0 (2024-04-24)
Platform: aarch64-apple-darwin20
Running under: macOS Ventura 13.2.1

Matrix products: default
BLAS:      /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
LAPACK:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib; LAPACK version 3.12.0

attached base packages:
[1] stats      graphics    grDevices   utils       datasets    methods     base

other attached packages:
[1] magick_2.8.3    patchwork_1.2.0  ggthemes_5.1.0  janitor_2.2.0
[5] here_1.0.1      lubridate_1.9.3 forcats_1.0.0  stringr_1.5.1
[9] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5    tidyverse_1.3.1
[13] vctrs_0.4.1     generics_0.1.2  ellipsis_0.3.2  grid_4.4.0
[17] gridExtra_2.3.3  lifecycle_0.3.0  pillar_1.9.0    lifecycle_0.3.0
[21] rappdirs_0.4.0  backports_1.4.1  assertthat_0.2.1  backports_1.4.1
[25] tools_4.4.0      R_4.4.0        compiler_4.4.0  compiler_4.4.0
```

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>

