Analysebericht: Wrangling und Datenvisualisierung

Bericht 1

VORNAME NACHNAME

2023-05-29

Inhaltsverzeichnis

```
knitr::opts_chunk$set(eval = T, # evaluate chunks
                      echo = T, # 'print code chunk?'
                      message = F, # 'print messages (e.g., warnings)?'
                      error = F, # stop when error encountered
                      warning = F) # don't print warnings
```

Beabsichtigte Lernergebnisse

- Das beabsichtigte Ziel dieses Berichts ist es, dass Sie nachweisen:

Kenntnisse und Fähigkeiten

- die Fähigkeit, R für die Datenverarbeitung und -visualisierung zu nutzen
- die Fähigkeit, Quarto und RStudio für reproduzierbare Berichte zu verwenden

Kommunikation

- die Fähigkeit, einen klar strukturierten und kohärenten Bericht zu erstellen

Bewertung

Die Bewertung erfolgt auf der Grundlage der Replikation der

- die Struktur des Dokuments (die HTML-Ausgabe)
- die Struktur und Genauigkeit des Codes (`.qmd`-Quellcode)
- die Ähnlichkeit zwischen den Originalplots und Ihren Plots

Tipps

Einige Tipps:

- Sie können auf Deutsch oder Englisch schreiben, je nachdem, was Sie bevorzugen
- Ich empfehle Ihnen dringend, Ihr Dokument *oft* zu rendern! Auf diese Weise können Sie, wenn ein Problem auftaucht, besser einschätzen, welcher Code es verursacht.

- Wenn Sie das Gefühl haben, dass Sie etwas langsam sind, versuchen Sie, sich auf das "große Ganze" zu konzentrieren und machen Sie sich keine Gedanken über all die Details in den Plots
  - Versuchen Sie zunächst, den Plot im Allgemeinen zu erstellen, und gehen Sie dann zu den Details über
  - Ich persönlich würde zuerst versuchen, die Handlung im Großen und Ganzen zu entwerfen, und mich dann um die Details kümmern

Tun Sie einfach Ihr Bestes!

Biondo, Soilemezidi, und Mancini (2022): Verarbeitung der Kongruenz der Zeitformen der Verben

Wir werden uns zum ersten Mal mit linguistischen Daten befassen. Sie müssen die Daten nicht vollständig verstehen, um die Aufgaben zu lösen, aber vielleicht finden Sie sie hilfreich. Ihre Hauptaufgabe besteht darin, vier Diagramme zur Untersuchung der Daten zu erstellen.

Unsere Daten stammen von Biondo, Soilemezidi, und Mancini (2022), die zwei Aufgaben umfassten:

1. ein Eye-Tracking-Leseexperiment
   - Die Teilnehmer lasen (spanische) Sätze mit Zeitformeln aus der Vergangenheit oder Zukunft (z. B. *ayer* 'gestern' oder *mañana* 'morgen') und Verbformen aus der Vergangenheit oder Zukunft (z. B. *compraron* 'gekauft' oder *comprarán* 'wird kaufen'). In Abbildung 1 finden Sie Beispielsätze in der Vergangenheit und in der Zukunft unter grammatikalischen ("match") und ungrammatikalischen ("mismatch") Bedingungen. Die + Augenbewegungen wurden aufgezeichnet, während die Teilnehmer die Sätze lasen
   - Die Haupthypothese: Wenn die Temporalphrasen nicht mit der Verbform übereinstimmen (z.B. 'gestern…kaufen' oder 'morgen…kaufen'), sollten *längere* Lesemaßnahmen in der Verbregion beobachtet werden.
   - Eine weitere Frage: Wird dieser Effekt sowohl für die Vergangenheit als auch für die Zukunft beobachtet?
2. eine Verb-'temporale Entscheidungsaufgabe'
   - In einer separaten Sitzung (d.h. nicht während des Eye-Tracking-Experiments) wurde den Teilnehmern ein isoliertes Verb präsentiert und sie wurden gebeten zu entscheiden, ob das Verb in der Vergangenheit oder in der Zukunft steht.
   - Genauigkeit und Reaktionszeit wurden aufgezeichnet.
   - Hauptforschungsfrage: Wie verhält sich die Bewertung der zeitlichen Referenz zwischen Verben in der Gegenwarts- und der Zukunftsform? Mit anderen Worten, wird das eine schneller erkannt als das andere?

Table 3
Experimental Material Addressing Question 1

Past	
Match	Mismatch
Gracias a la beca, el año pasado los investigadores progresaron en sus estudios sobre la polución. "Thanks to the scholarship, last year the researchers made progress on their studies on pollution"	Gracias a la beca, el año pasado los investigadores progresarán en sus estudios sobre la polución. "Thanks to the scholarship, last year the researchers will make progress on their studies on pollution"
Future	
Match	Mismatch
Gracias a la beca, el próximo año los investigadores progresarán en sus estudios sobre la polución. "Thanks to the scholarship, next year the researchers will make progress on their studies on pollution"	Gracias a la beca, el próximo año los investigadores progresaron en sus estudios sobre la polución. "Thanks to the scholarship, next year the researchers made progress on their studies on pollution"

Note. The underlined terms represent the target region.

Abbildung 1: Beispielsätze aus Biondo, Soilemezidi, und Mancini (2022)

Variablen

Unsere **Messvariablen** (d. h. abhängige Variablen) sind:

- die Eye-Tracking-Lesemessung ***total reading time*** in der Verbregion (die Gesamtzeit, die beim Lesen des Satzes auf das Verb geschaut wird)
- die ***Reaktionszeiten*** (aus der zeitlichen Entscheidungsaufgabe (*temporal decision task*))
- die ***Genauigkeit*** (aus der zeitlichen Entscheidungsaufgabe (*temporal decision task*))

Unsere **Prädiktorvariablen** (d. h. unabhängige Variablen) sind:

- die ***Verbform*** (*tense*) (Vergangenheit oder Zukunft)
- ***Grammatikalität*** (ob es grammatikalisch oder ungrammatikalisch war, basierend auf dem vorangehenden Temporaladverb)

Vorhersagen

In der Regel gehen wir davon aus, dass längere Reaktions- und/oder Lesezeiten (wie die Gesamtlesezeit) auf *Verarbeitungsschwierigkeiten*, d. h. Schwierigkeiten bei der Sprachverarbeitung, hinweisen. Die Hypothese war daher, dass längere Gesamtlesezeiten für die *Verb*-Region in ungrammatischen Sätzen vorhanden sein würden. Bei den Reaktionszeiten geht es um die Frage, ob die Zeitformen (Vergangenheit/Zukunft) bei einer Kategorisierungsaufgabe ähnliche Reaktionszeiten hervorrufen oder ob eine Zeitform länger für die Kategorisierung braucht als die andere. All diese Fragen werden wir heute mit den Daten nicht beantworten können.

1 Einrichten

1.1 YAML

- Erstellen Sie ein neues Quarto-Dokument
- Stellen Sie sicher, dass Ihre YAML enthält:
  - einen passenden Titel
  - Ihren Vor- und Nachnamen (`author`)
  - das Datum
  - das Inhaltsverzeichnis
  - unter `format: html…`, `self-contained: true` einfügen
- Ihr YAML sollte etwa so aussehen

```yaml
---
title: "Summary assignment: wrangling and data viz"
subtitle: "In-class assignment 1"
author: "YOUR NAME HERE"
institute: Humboldt-Universität zu Berlin
lang: de
date: "`r Sys.Date()`"
format:
  html:
    toc: true
    number-sections: true
    self-contained: true
---
```

1.2 Struktur des Dokuments

- Sie werden Ihr Quarto-Skript und die HTML-Ausgabe Ihres Skripts einreichen.
- Um die Lesbarkeit des gerenderten Dokuments zu verbessern, sollten Sie das Dokument durch Überschriften und Zwischenüberschriften strukturieren.
- Verwenden Sie schriftlichen Text, wo es notwendig ist, um Fragen zu beantworten/ Prozesse zu beschreiben
- Verwenden Sie bei Bedarf Code-Bausteine
- eine gute Strategie für die Wahl einer Überschrift: Wenn dieses Dokument einen neuen Abschnitt enthält, sollten Sie auch einen neuen Abschnitt erstellen

Tabelle 1: Variable names and descriptions for dataset Biondo_etal_2021.csv

	Description
Tense_type	Tense (past/future)
subj	Participant ID
Item.num	Item number
verb	verb
acc	accuracy (0 = wrong, 1 = correct)
RT	Reaction time (milliseconds)
totalTime	total reading time (milliseconds)
gramm	grammatical (0 = no, 1 = yes)

1.3 Pakete

- werden Sie die Pakete benötigen:
  - `tidyverse`
  - `here`
  - `patchwork`
  - `ggthemes`

Laden Sie sie mit dem Paket `pacman` (Funktion `p_load()`).

```
pacman::p_load(tidyverse,
               here,
               ggthemes,
               patchwork)
```

1.4 Datenbeschreibung

Laden Sie die folgenden Daten herunter und speichern Sie sie im Ordner `daten`: `Biondo_etal_2021.csv`.

Unsere Variablen sind in Tabelle 1 beschrieben.

2 Datenexploration

2.1 Data-Wrangling

1. Um die Daten zu laden, kopieren Sie den nachstehenden Code.

```
df_resp <- read_csv(here("daten", "Biondo_etal_2021.csv"))
```

1. Um die Daten zu laden, kopieren Sie den nachstehenden Code. 2. Ergänzen Sie die obige Codezeile (mit einer Pipe) so, dass Sie auch:

   - Benennen Sie die Variablen um:
     - `Tense_type` als `tense`
     - `Item.num` als `item`
     - `RT` als `rt`
     - `totalTime` als `tt`
   - Verschieben Sie `tense` vor `verb`, und `gramm` nach `verb`.
   - Ordne die Zeilen nach `subj` und nach `item`

```
df_resp <- read_csv(here("daten", "Biondo_etal_2021.csv")) %>%
  rename(tense = Tense_type,
         item = Item.num,
         rt = RT,
         tt = totalTime) %>%
  relocate(tense, .before = verb) %>%
  relocate(gramm, .after = verb) %>%
  arrange(subj, item)
```

Hinweis: Der Kopf Ihrer Daten sollte wie folgt aussehen:

```
head(df_resp)
```

```
# A tibble: 6 x 8
   subj  item tense  verb           gramm   acc    rt    tt
  <dbl> <dbl> <chr>  <chr>          <dbl> <dbl> <dbl> <dbl>
1     1     1 future representarán      1     1  840.  1596
2     1     2 future alzarán            1     1 1310.   648
3     1     3 future centrarán          1     1  700.   841
4     1     4 future coleccionarán      1     1  650.  1337
5     1     5 future complementarán     1     1  580.  1400
6     1     6 future despilfarrarán     1     1  610.  1649
```

2.2 Erkunden die Daten

2. Untersuchen Sie den Datensatz mithilfe der entsprechenden Funktion(en). Nehmen Sie nur das in den Bericht auf, was Sie für die *beste* Zusammenfassung halten (d. h., Sie können mit der Datenauswertung herumspielen, aber nehmen Sie nur das auf, was Sie für eine prägnante und informative Zusammenfassung halten).

```
summary(df_resp)
```

```
      subj            item           tense               verb
 Min.   : 1.00   Min.   :  1.00   Length:5760        Length:5760
 1st Qu.:22.75   1st Qu.: 24.75   Class :character   Class :character
 Median :37.50   Median : 51.50   Mode  :character   Mode  :character
 Mean   :37.97   Mean   : 50.86
 3rd Qu.:54.25   3rd Qu.: 77.25
 Max.   :74.00   Max.   :101.00

     gramm            acc               rt               tt
 Min.   :0.0000   Min.   :0.0000   Min.   : 180.3   Min.   :  90.0
 1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.: 820.4   1st Qu.: 324.0
 Median :1.0000   Median :1.0000   Median :1045.3   Median : 489.5
 Mean   :0.6667   Mean   :0.9538   Mean   :1250.4   Mean   : 600.6
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1430.3   3rd Qu.: 736.0
 Max.   :1.0000   Max.   :1.0000   Max.   :7661.4   Max.   :3936.0
                                                    NA's   :72
```

```
glimpse(df_resp)
```

```
Rows: 5,760
Columns: 8
$ subj  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ item  <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 1~
$ tense <chr> "future", "future", "future", "future", "future", "future", "fut~
$ verb  <chr> "representarán", "alzarán", "centrarán", "coleccionarán", "compl~
$ gramm <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ acc   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
$ rt    <dbl> 840.1917, 1310.1809, 700.2674, 650.1856, 580.2159, 610.2178, 760~
$ tt    <dbl> 1596, 648, 841, 1337, 1400, 1649, 1504, 718, 1453, 500, 778, 160~
```

2.3 Variablen

1. Erstellen Sie eine Aufzählung der Variablen, in der Sie angeben, um welchen *Typ* von Variable es sich jeweils handelt (z. B. numerisch, kategorisch usw.). Denken Sie daran, dass Zahlen nicht immer numerisch sind. Setzen Sie die Variablennamen kursiv.

- *Variablenname*: Variablentyp
- …

2. Untersuchen Sie das Dichteplot (Abbildung 2 A) und das Histogramm (Abbildung 2 B) unten. Beobachten Sie, was auf der x-Achse aufgetragen wird, und beantworten Sie die folgenden Fragen:

    b. Was zeigen die beiden Diagramme (benenne, was entlang der x- und y-Achse aufgetragen ist; wofür stehen die verschiedenen Farben?).

c. Zeigen die beiden Diagramme die gleichen Daten?
d. Welches `geom` wird benötigt, um ein Dichte-Diagramm zu erstellen? Ein Histogramm? (e.g., `geom_barplot()` wird zur Erstellung von Balkendiagrammen verwendet)

```
fig_density <- df_resp %>%
  ggplot(aes(x = rt, fill = as_factor(acc), colour = as_factor(acc))) +
  geom_density(alpha = 0.4) +
  labs(title = "Reaction times per accuracy level",
       x = "Reaction times (ms)",
       y = "Density",
       fill = "Accuracy",
       colour = "Accuracy") +
  theme_minimal()
```

```
fig_histogram <- df_resp %>%
  ggplot(aes(x = rt, fill = as_factor(acc), colour = as_factor(acc))) +
  facet_grid(.~as_factor(acc)) +
  geom_histogram(alpha = 0.4) +
  labs(title = "Reaction times per accuracy level",
       x = "Reaction times (ms)",
       y = "Count",
       fill = "Accuracy",
       colour = "Accuracy") +
  theme_bw()
```

```
((plot_spacer() + fig_density + plot_spacer() + plot_layout(nrow = 1, widths = c(.15,.7
   fig_histogram) +
  plot_layout(nrow = 2) +
  plot_annotation(tag_levels = "A")
```

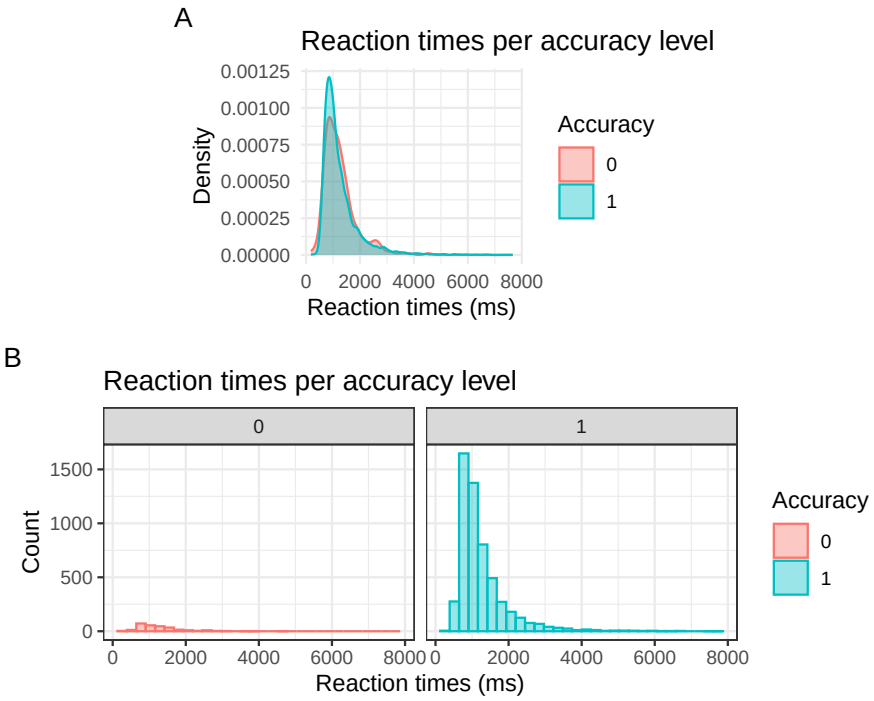

Abbildung 2: Dichte- und Histogrammdiagramme

3. Beschreiben Sie die Verteilung der Reaktionszeiten auf der Grundlage dieser Diagramme. Wie hoch sind die ungefähren Reaktionszeiten der richtigen Antworten? Für ungenaue Antworten? Gibt es mehr genaue oder ungenaue Antworten?

3 Datenvisualisierung

Replizieren *und beschreiben* Sie die folgenden Diagramme.

3.1 Balkenplot

1. Reproduzieren Sie Abbildung 3 (Genauigkeit durch Anspannung). Wir haben nicht besprochen, wie man die Balken nebeneinander druckt. ***Tipp***: Dies nennt man einen *gruppierten* Balkenplot (DE: grouped barplot). Möglicherweise müssen Sie "grouped barplot ggplot2" googeln, um herauszufinden, wie man einen solchen Barplot erstellt.

```
df_resp %>%
  ggplot(aes(x = as_factor(tense), fill = as_factor(acc))) +
  geom_bar(position="dodge") +
  labs(title = "Accuracy by tense",
       x = "Tense",
```

```
      y = "Count",
      fill = "Accuracy",
      colour = "Accuracy") +
  scale_fill_colorblind() +
theme_bw()
```

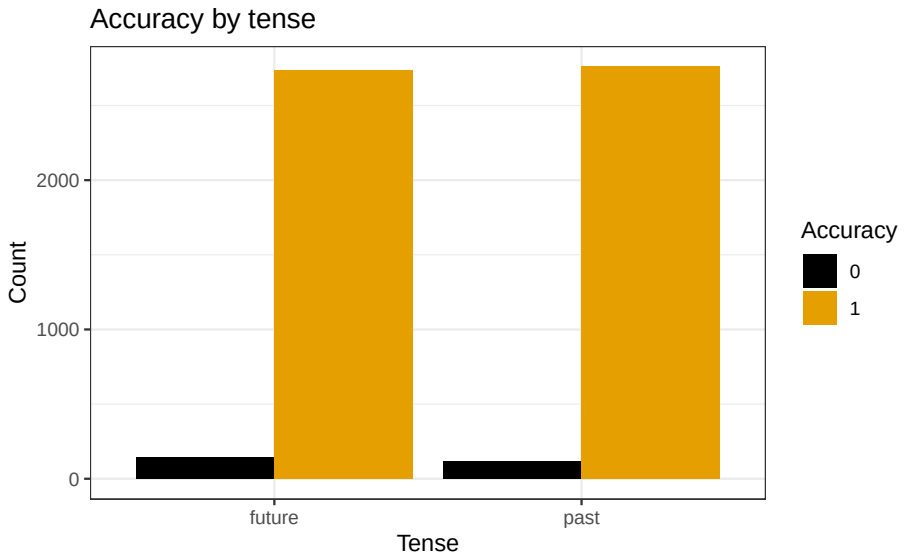

Abbildung 3: Balkenplot

Beschreiben Sie die Handlung, z. B. ob die Antworten in den einzelnen Zeitformen genauer oder ungenauer waren? Waren sie zwischen den Zeitformen ähnlich?

3.2 Punktwolke

2. Reproduzieren Sie Abbildung 4 und beantworten Sie diese Frage: Gibt es einen Trend zwischen Reaktionszeiten und Gesamtlesezeiten? Beschreiben Sie, was Sie sehen.

```
df_resp %>%
  ggplot(aes(x = tt, y = rt)) +
  geom_point(position = position_jitterdodge(.5), alpha = .5,
             aes(colour = tense, shape = tense)) +
  labs(title = "Reaction time by total reading time at the verb region",
       x = "Total reading time (ms)",
       y = "Reaction time (ms)",
       shape = "Tense",
       colour = "Tense") +
    scale_fill_colorblind() +
```

```
  geom_smooth(method="lm") +
  theme_bw()
```

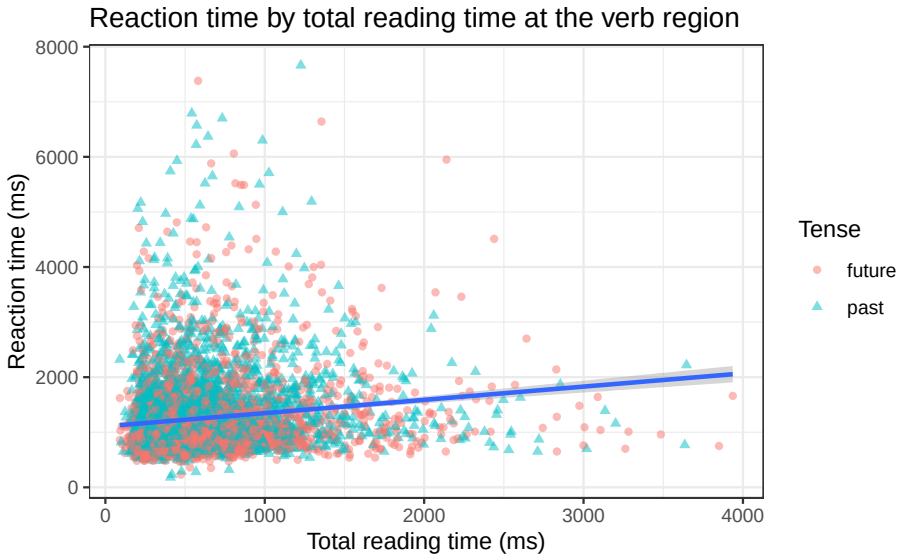

Abbildung 4: Punktwolke

4 Interpretieren eines neuen Diagrammtyps

Eine wichtige Fähigkeit besteht darin, zu lernen, wie man Daten in einer Darstellung interpretiert. In Abbildung 5 sehen Sie einen Boxplot (auch Box-and-Whisker-Plot genannt), den wir bisher noch nicht gesehen haben. Beantworten Sie die folgenden Fragen zu diesem Diagramm:

1. Was ist auf der x-Achse aufgetragen? Handelt es sich um einen numerischen oder kategorialen Faktor?
2. Was ist auf der y-Achse eingezeichnet? Handelt es sich um einen numerischen oder kategorialen Faktor?
3. Kannst du den Namen des `geom` erraten, das einen Boxplot erzeugt? Denke über den Namen eines `geom` für ein Balkendiagramm nach, zum Beispiel.
4. Was glaubst du, was die Punkte im Diagramm darstellen? (Es ist okay, sich zu irren.)
5. Was glaubst du, stellt das Kästchen auf dem Diagramm dar? (Es ist in Ordnung, sich zu irren.)
6. Was glaubst du, stellt die dicke Linie in der Mitte des Kästchens dar? (Es ist in Ordnung, sich zu irren.)

```
df_resp %>%
  ggplot(aes(x = as_factor(gramm), y = tt,
```

```
            colour = as_factor(gramm))) +
facet_grid(.~as_factor(tense)) +
geom_boxplot() +
labs(title = "Total reading time (verb region) by tense and grammaticality",
     y = "Total reading time (ms)",
     x = "Grammaticality") +
  scale_fill_colorblind() +
theme_bw() +
theme(legend.position = "none")
```

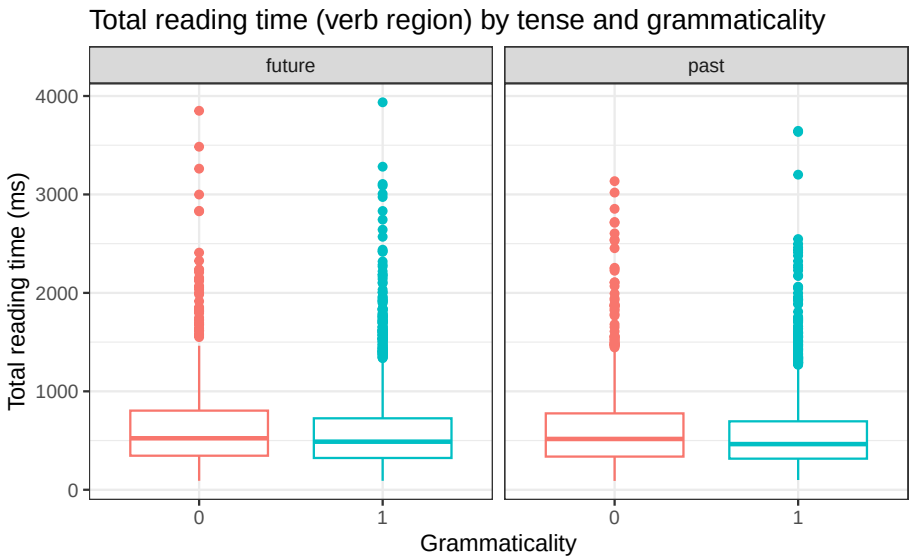

Abbildung 5: Boxplot

5 Session Info einbeziehen

Am Ende Ihres Berichts:

1. Erstellen Sie einen neuen Abschnitt (`# Session Info`)
2. Fügen Sie den folgenden Text als Inline-Text ein:

   Hergestellt mit `` `r R.version.string` `` (`` `r R.version$nickname` ``).

3. Fügen Sie das Folgende in einen Codeabschnitt ein:

```
sessionInfo()
```

Wenn Sie nun Ihr Dokument rendern, sollte eine Zusammenfassung der geladenen Pakete und ihrer Versionen angezeigt werden.

6 Bericht einreichen

Wenn Sie fertig sind, laden Sie Ihr `.qmd`-Skript *und* die gerenderte Ausgabe im HTML-Format in Moodle hoch.

6.1 Früher fertig?

Wenn Sie zu früh fertig waren oder die Aufgaben zu einfach fanden, versuchen Sie, die obigen Dichte- und Histogrammdiagramme zu reproduzieren. ***Tipp***: Für das Histogramm müssen Sie googeln, um herauszufinden, wie Sie die Darstellung in zwei Boxen aufteilen können (die Boxen werden `facet` genannt)

Session Info

Hergestellt mit R version 4.3.0 (2023-04-21) (Already Tomorrow) und RStudioversion 2023.3.0.386 (Cherry Blossom).

```
sessionInfo()
```

```
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;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: Europe/Berlin
tzcode source: internal

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

other attached packages:
 [1] patchwork_1.1.2 ggthemes_4.2.4  here_1.0.1      lubridate_1.9.2
 [5] forcats_1.0.0   stringr_1.5.0   dplyr_1.1.2     purrr_1.0.1
 [9] readr_2.1.4     tidyr_1.3.0     tibble_3.2.1    ggplot2_3.4.2
[13] tidyverse_2.0.0

loaded via a namespace (and not attached):
```

```
 [1] utf8_1.2.3           generics_0.1.3       xml2_1.3.4
 [4] lattice_0.21-8       stringi_1.7.12       hms_1.1.3
 [7] digest_0.6.31        magrittr_2.0.3       evaluate_0.21
[10] grid_4.3.0           timechange_0.2.0     fastmap_1.1.1
[13] Matrix_1.5-4         rprojroot_2.0.3      jsonlite_1.8.4
[16] mgcv_1.8-42          httr_1.4.6           rvest_1.0.3
[19] fansi_1.0.4          viridisLite_0.4.2    scales_1.2.1
[22] cli_3.6.1            rlang_1.1.1          crayon_1.5.2
[25] splines_4.3.0        bit64_4.0.5          munsell_0.5.0
[28] withr_2.5.0          yaml_2.3.7           tools_4.3.0
[31] parallel_4.3.0       tzdb_0.4.0           colorspace_2.1-0
[34] webshot_0.5.4        pacman_0.5.1         kableExtra_1.3.4.9000
[37] vctrs_0.6.2          R6_2.5.1             lifecycle_1.0.3
[40] bit_4.0.5            vroom_1.6.3          pkgconfig_2.0.3
[43] pillar_1.9.0         gtable_0.3.3         glue_1.6.2
[46] systemfonts_1.0.4    xfun_0.39            tidyselect_1.2.0
[49] rstudioapi_0.14      knitr_1.42           farver_2.1.1
[52] nlme_3.1-162         htmltools_0.5.5      labeling_0.4.2
[55] svglite_2.1.1        rmarkdown_2.21       compiler_4.3.0
```

Literaturverzeichnis

Biondo, Nicoletta, Marielena Soilemezidi, und Simona Mancini. 2022. „Yesterday Is History, Tomorrow Is a Mystery: An Eye-Tracking Investigation of the Processing of Past and Future Time Reference During Sentence Reading.“ *Journal of Experimental Psychology: Learning, Memory, and Cognition* 48 (7): 1001–18. https://doi.org/10.1037/xlm0001053.