

Bericht 1

Daniela Palleschi

Di. den 28.05.2024

Inhaltsverzeichnis

1	Beschreibung	1
2	Einrichtung	1
2.1	Quarto	1
2.2	Pakete	2
2.3	Daten	2
3	Data wrangling	3
3.1	<code>filter()</code>	3
3.2	<code>mutate()</code>	3
3.3	Fehlersuche	4
4	Datenvisualisierung	4
4.1	Streudiagramm	4
4.2	Facetten	4
4.3	Reproduzieren eines Plots	5
5	Interpretation	5

1 Beschreibung

Dieser Bericht dient dazu, das bisher Gelernte zu wiederholen und zu festigen. Ihre Aufgaben umfassen das Laden von Paketen und Daten sowie eine leichte Datenverarbeitung (Abschnitt 3). Außerdem werden Sie 4 Diagramme erstellen (Abschnitt 4) und eine kurze Interpretation zu einem der Diagramme schreiben (Abschnitt 5). Ein Tipp: Ich empfehle Ihnen, Ihr Dokument häufig zu rendern, um eventuelle Fehler frühzeitig zu erkennen.

Bitte reichen Sie sowohl das gerenderte PDF-Dokument als auch das Quarto-Skript ein.

2 Einrichtung

2.1 Quarto

Öffnen Sie ein neues Quarto-Skript und speichern Sie es als `bericht1_nachname_vorname.qmd`. Nehmen Sie die folgenden Änderungen an der YAML vor:

- PDF als Format
- einen aussagekräftigen Titel
- Ihren Namen als **Autor**
- ein Inhaltsverzeichnis (`toc: true`)

Sie können auch die YAML aus einem früheren Skript kopieren, wenn dieses die gleichen Einstellungen hat.

Achten Sie darauf, Code Chunks, Prosa und Überschriften zu verwenden, um Ihre Aufgaben angemessen zu dokumentieren. Eine gute Faustregel ist, für jede (Unter-)Überschrift in diesem Dokument eine Überschrift hinzuzufügen.

2.2 Pakete

Laden Sie die Pakete `tidyverse` und `languageR` ein.

```
```{r}
Pakete laden...
```
```

2.3 Daten

Der Datensatz `durationsGe` aus dem `languageR`-Paket (Baayen & Shafaei-Bajestan, 2019) enthält Dauermessungen zur niederländischen Vorsilbe *ge*. Eine Beschreibung aller Variablen des Datensatzes findet sich in Tabelle 1. Ihre Aufgabe ist es:

1. Speichern Sie den Datensatz als Objekt `df_ge` in Ihrer Umgebung (dies kann auf die gleiche Weise geschehen wie bei allen Datensätzen, die wir bisher verwendet haben)
2. Drucken Sie die ersten 10 Zeilen des Datensatzes mit der Funktion `head()` aus.

```
```{r}
1. Daten laden
```
```

```
```{r}
2. erste 10 Zeilen
```
```

Tabelle 1: ‘Data dictionary’ für ‘durationsGe’ aus dem languageR-Paket

| Variable | Beschreibung |
|------------------|--|
| Word | ein Faktor mit den Wörtern als Ebenen |
| Frequency | ein numerischer Vektor mit der absoluten Häufigkeit des Wortes im Spoken Dutch Corpus |
| Speaker | ein numerischer Vektor mit der absoluten Häufigkeit des Wortes im Spoken Dutch Corpus |
| Sex | ein Faktor mit den Lautsprechern als Ebenen |
| YearOfBirth | ein numerischer Vektor mit Geburtsjahren |
| DurationOfPrefix | ein numerischer Vektor mit der Dauer des Präfixes -ge in Sekunden. |
| SpeechRate | ein numerischer Vektor, der die Sprechgeschwindigkeit in Anzahl der Silben pro Sekunde kodiert |
| NumberSegments | ein numerischer Vektor, der die Sprechgeschwindigkeit in Anzahl der Silben pro Sekunde kodiert |

3 Data wrangling

Hier werden Sie die `dplyr`-Verben aus Woche 4 verwenden. Denken Sie daran, dass Sie den Zuweisungsoperator (`<-`) nur verwenden müssen, wenn Sie die Änderungen, die Sie vornehmen, als Objekt in der Umgebung speichern wollen. Wenn Sie diese Änderungen nur ausdrucken wollen, brauchen Sie den Zuweisungsoperator nicht.

3.1 `filter()`

Drucken (aber nicht in Ihrer Umgebung speichern) Sie die Zeilen von `df_ge`, in denen `SpeechRate` über 9 liegt, nur mit den Spalten `word`, `speaker` und `SpeechRate`. Es sollten 5 Zeilen sein.

```
```{r}
subsetting
```
```

3.2 mutate()

Fügen Sie eine neue Variable hinzu, `duration_ms`, die `DauerVonPräfix` multipliziert mit 1000 (`DurationOfPrefix*1000`) entspricht. Dies entspricht der Dauer von *ge* in Millisekunden, statt in Sekunden. Stellen Sie sicher, dass Sie diese neue Variable in Ihrem Datenrahmen speichern (Hinweis: Sie müssen den Zuweisungsoperator `<-` und das `dplyr`-Verb `mutate()` verwenden).

```
```{r}
mutate(): duration_ms
```
```

3.3 Fehlersuche

Warum läuft dieser Code nicht? Es gibt zwei Probleme mit dem Code, identifizieren und beheben Sie sie.

```
# Troubleshooting
df_ge |>
  filter(YearOfBirth == 1978) +
  select(Frequency, word)
```

```
```{r}
Lösung
```
```

4 Datenvisualisierung

Verwenden Sie für alle Diagramme `labs(title = "...")`, um entsprechende Diagrammtitel hinzuzufügen.

Optional: Ändern Sie die `x` und `y` Achsenbeschriftungen, wenn Sie wollen, mit `labs(x = "...", y = "...")`. Vielleicht möchten Sie auch ein Thema hinzufügen (z.B. `theme_minimal()`).

4.1 Streudiagramm

Erstellen Sie ein Streudiagramm (`geom_point()`) mit `SpeechRate` (x-Achse) und `DurationOfPrefix` (y-Achse), mit `YearOfBirth` als Farbe (`colour`). Ändern Sie die Einstellungen für den Codechunk so, dass das Diagramm beim Rendern des Skripts *nicht* gedruckt wird, der Code aber schon. Tipp: Sie müssen `#| eval:` verwenden.

```
```{r}
Streudiagramm + #| eval:
```
```

4.2 Facetten

Fügen Sie Facetten für `Sex` hinzu (denken Sie daran, die Tilde `~` einzufügen). Ändern Sie die Code-Chunk-Einstellungen so, dass die Darstellung gedruckt wird, wenn das Skript gerendert wird, aber der Code nicht (Sie benötigen `echo` anstelle von `eval`).

```
```{r}
Streudiagramm + Facetten + #| echo:
```
```

4.3 Reproduzieren eines Plots

Reproduzieren Sie die Abbildung 1 (es muss keine exakte Kopie sein, aber kommen Sie ihr so nahe wie möglich). Stellen Sie sicher, dass sowohl der Code als auch die Darstellung beim Rendern gedruckt werden. Hinweis: Sie müssen `filter()` sowohl für `Frequency` als auch für `Sex` verwenden. Ich würde mich darauf konzentrieren, zuerst das Diagramm zu erstellen und dann zu versuchen, die Daten zu filtern.

5 Interpretation

Beschreiben Sie die Beziehung zwischen den beiden Variablen, die Sie in Abbildung 1 sehen.

Baayen, R. H., & Shafaei-Bajestan, E. (2019). *languageR: Analyzing Linguistic Data: A Practical Introduction to Statistics*. <https://CRAN.R-project.org/package=languageR>

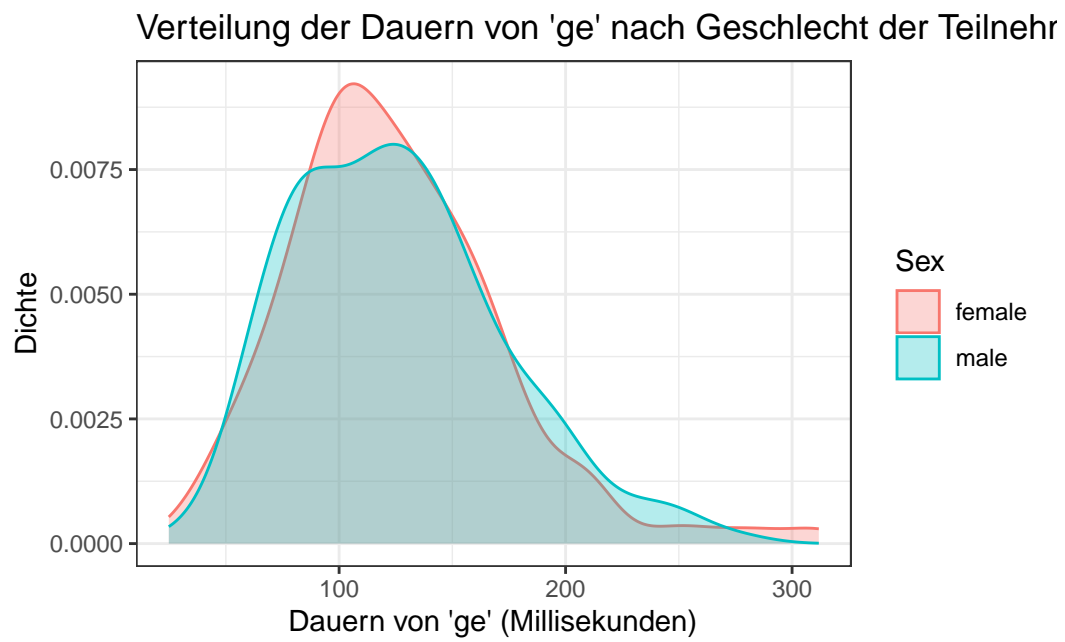


Abbildung 1: Eine zu reproduzierende Abbildung