Computergestützte Datenanalyse: DATA-Übung mit R

Tag 3 – 24.07.2025

UNSER PLAN

- **Tag 1**
  - Einführung in R und RStudio
  - „Basics“:
  - Coding Konventionen
  - Objekte, Datenimport & Co
- **Tag 2**
  - Skalenniveau
  - Troubleshooting
  - Datenaufbereitung
  - Datenvisualisierung
  - Deskriptive Statistik
- **Tag 3**
  - Bivariate Analyse
- **Tag 4**
  - Inferenzstatistik
  - Abschluss

Genereller Ablauf

- Vier Tage geblockt
- Mischung aus Input- und Übungssessions
- Anwesenheitsabfrage alle 90 Minuten

Heute

- Zwei 15 Minuten Pause
- Eine Mittagspause

REFRESHER

- Was sind Funktionen?
- Was sind die vier Skalenniveaus?
- Welche Datentypen/-klassen gibt es?
- Was tun, wenn Error Meldungen auftauchen?
- Was ist der Unterschied zwischen Subsetting und Rekodieren?

  Welche Packages, Funktionen?
- Warum visualisieren wir Daten? Welches Package hilft dabei?

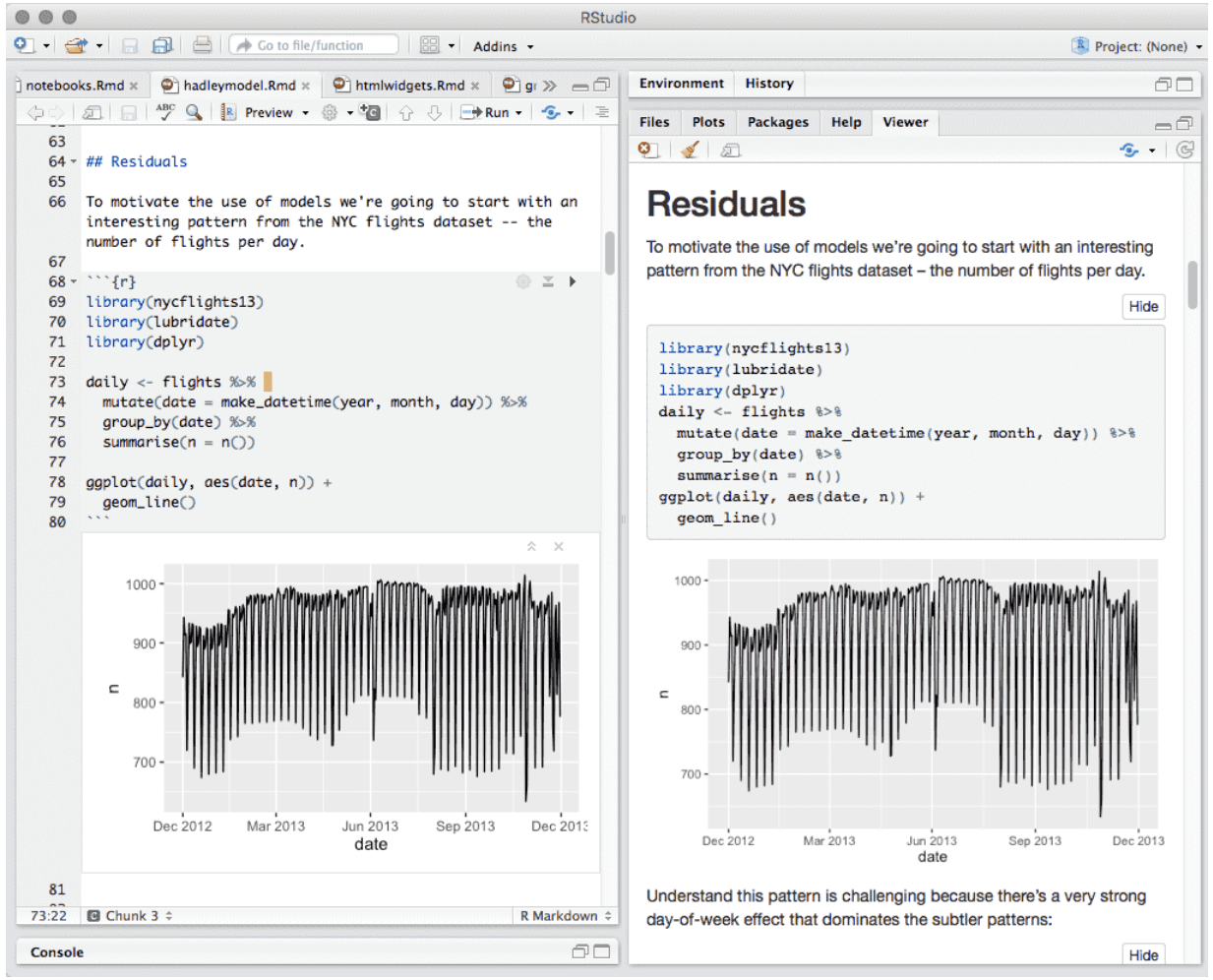
Welche Datenanalysemethoden sind Ihnen bekannt?

Welche Datenanalyse würden Sie gerne mit R durchführen, wenn Ihren Daten und Programmierkenntnissen keine Grenzen gesetzt sind?

5 Minuten untereinander diskutieren, kurz im Plenum vorstellen

R Markdown

- Kombination aus Code und Text
- Gut zum…
  - Teilen
  - Veranschaulichen



Getting started

1. Rmarkdown file erstellen und einstellen
2. Packages installieren und/ oder aktivieren: rio und tidyverse
3. Import der ALLBUS 2018 Daten
4. In einem **neuem** Dataframe werden die Variablen pt01 bis pt20 gespeichert. Nutzt dafür den Befehl select() und recherchiert, wie das funktioniert
5. Um welche Variablen handelt es sich? (s. im Variablen Report nach und kommentiert euer Rmd)
6. Berechnet Mittelwert und Standardabweichung für die Variablen pt01, pt02 und pt03, in Textform reporten! Beispiel: pt01 hat einen *M*=X und *SD*=Y.
7. Untersucht die Verteilungen, ergo absolute und relative Häufigkeiten für pt03, in Textform reporten!
8. Erstellt einen barplot für pt03 mit Beschriftung

Exkurs: Boxplot erstellen

Was ist ein Boxplot?

```
ggplot(data = ?, aes(y = pt01)) +
  geom_boxplot() +
  labs(title = "?",
       y = "?")
```

Bivariate Analysen

Kreuztabellen und Zusammenhangsmaße

- Typische Frage: Gibt es Unterschiede in der Verteilung zwischen zwei Variablen?

- Anzahl der Ausprägungen beachten -> siehe Kapitel zur Häufigkeitsauszählung

Aufgabe:

```
tabelle <- table(df$pt01, df$pt02)
```

Was ist zu sehen?

Bivariate Analyse mittels Kreuztabelle: Test auf statistische Unabhängigkeit

- Untersucht den statistischen Zusammenhang zwischen zwei Variablen
  - Gibt es einen überzufälligen Zusammenhang?
  - Falls ein überzufälliger Zusammenhang besteht: Wie stark ist der Zusammenhang?
  - Wie können wir den Zusammenhang interpretieren?
- Gängiges Werkzeug zur Analyse von Daten auf nominalem oder ordinalem Messniveau
  - Zum Beispiel: Chi-Quadrat, Cramer‘s V oder Phi (bei 2x2-Tabellen sind letztere identisch)

Zusammenhangsmaß

- Gibt an, wie stark der Zusammenhang zwischen zwei Variablen ist

- Das Zusammenhangsmaß ist abhängig vom Skalenniveau:
  - Zwei dichotome Variablen: phi (ϕ)
  - Eine dichotome und eine nominale Variable, zwei nominale Variablen, eine ordinale und eine dichotome oder nominale Variable: Cramer's V
  - Zwei mindestens ordinalskalierte Variablen oder eine ordinale und eine mindestens intervallskalierte Variable: Spearman's rho (ρ) oder Kendall's tau-b (τ -b) oder tau-c(τ -c)
  - Zwei mindestens intervallskalierte Variablen: Pearson's r
- Zusammenhangsmaße liegen zwischen 0 und 1 beziehungsweise zwischen -1 und +1
  - > 0.2 interpretierbarer Zusammenhang
  - > 0.5 starker Zusammenhang
  - *1 perfekter Zusammenhang*

Chi-Quadrat (χ^2)

- Test auf statistische Unabhängigkeit
- Je größer χ^2, desto stärker der Zusammenhang von zwei Variablen (0=kein Zusammenhang)

Aufgabe

Kreuztabelle erstellen:

tabelle <- table(df$variable, df$variable)

Chi^2-Test durchführen mithilfe der Kreuztabelle

chisq.test(tabelle)

Und was sagt uns das Ergebnis?

Chi-Quadrat (χ^2)

- Test auf statistische Unabhängigkeit
- Je größer χ^2, desto stärker der Zusammenhang von zwei Variablen (0=kein Zusammenhang)
- Problem bei der **Interpretation**:
  - Kann sehr große Werte annehmen
  - Zusammenhang hängt von Tabellenformat und Fallzahl ab

- Lösung: Normierung

Cramers V

- Normiert einen χ^2-Wert von 0 bis 1
  - Unempfindlich gegenüber Tabellenformat und Fallzahl
- 0 bedeutet kein Zusammenhang
- 1 bedeutet perfekter Zusammenhang

Aufgabe

chisq.test(tabelle)

install.packages("vcd")
library(vcd)

assocstats(tabelle)

Schreibt eine kurze Interpretation

Zur Erinnerung:

0 bedeutet kein Zusammenhang
1 bedeutet perfekter Zusammenhang