Block 1: Grundlagen Data Science

Topic 3: Praktische Pipeline & EDA

R: Die Arbeitsumgebung

Warum R?

- Spezialisiert auf Statistik und Datenanalyse
- Exzellente Visualisierungsmöglichkeiten
- Open Source und kostenlos
- Große Community und viele Pakete

RStudio IDE

- Integrierte Entwicklungsumgebung
- Editor, Console, Environment, Plots
- Projektmanagement

R-Projekt-Struktur

- `projekt.Rproj` – Projektdatei
- `data/` – Rohdaten
- `scripts/` – R-Skripte
- `output/` – Ergebnisse, Plots
- `README.md` – Dokumentation

Vorteil: Alles an einem Ort, reproduzierbar

R-Projekte ermöglichen strukturierte, reproduzierbare Analysen

Relative Pfade

Absolute Pfade (vermeiden!)

- `C:/Users/Anna/projekt/data.csv`
- Funktioniert nur auf einem PC
- Code nicht portabel
- Kollaboration erschwert

Relative Pfade (gut!)

- `data/data.csv`
- Relativ zum Projektordner
- Reproduzierbarkeit gewährleistet
- Code teilbar

Best Practice: `here::here()`

- Findet automatisch Projekt-Root
- `here::here("data", "data.csv")`
- Funktioniert überall

Beispiel:

```
library(here)
df <- read_csv(
  here("data", "sales.csv")
)
```

Relative Pfade garantieren Reproduzierbarkeit über verschiedene Systeme hinweg

Das Tidyverse

Core Packages

- `dplyr` – Datenmanipulation
- `ggplot2` – Visualisierung
- `tidyr` – Daten aufräumen
- `readr` – Daten einlesen
- `tibble` – Moderne Data Frames
- `purrr` – Funktionale Programmierung

Pipe-Operator %>%

- Verketten von Operationen
- Lesbarer Code
- Von links nach rechts

Beispiel-Pipeline:

```
library(tidyverse)

df %>%
  filter(sales > 1000) %>%
  mutate(
    profit = sales - costs
  ) %>%
  group_by(region) %>%
  summarise(
    avg_profit = mean(profit)
  ) %>%
  arrange(desc(avg_profit))
```

Lesbarkeit: Jeder Schritt klar erkennbar

Das Tidyverse bietet konsistente, lesbare Syntax für moderne Datenanalyse

Explorative Datenanalyse (EDA)

Ziele der EDA

- Daten verstehen und kennenlernen
- Verteilungen erkunden
- Muster und Zusammenhänge finden
- Ausreißer identifizieren
- Hypothesen generieren

EDA Checklist

- `dim(df)` – Dimensionen
- `str(df)` – Struktur
- `summary(df)` – Statistiken
- `head(df)` – Erste Zeilen
- Fehlende Werte prüfen

Visualisierungstypen

- **Univariat:**
  - Histogramm
  - Boxplot
  - Density Plot
- **Bivariat:**
  - Scatterplot
  - Korrelationsmatrix
- **Multivariat:**
  - Faceted Plots
  - Pairplots

Regel: Erst verstehen, dann modellieren!

EDA ist der erste kritische Schritt vor jeder Modellierung