

Sommersemester 2026

Datenmanagement & -analyse

Prof. Dr. Christoph M. Flath

Lehrstuhl für Wirtschaftsinformatik und Business Analytics

Julius-Maximilians-Universität Würzburg

► 1 Motivation & EDA-Ziele

2 Univariate Analyse

3 Bivariate Analyse

4 Visualisierung

5 SQL für EDA

6 Praktische EDA

7 Zusammenfassung

Letzte Sessions:

- JOINS – Tabellen verknüpfen
- Subqueries – Abfragen verschachteln
- CTEs – Lesbare komplexe Abfragen
- Views – Gespeicherte Abfragen
- Transaktionen – ACID

Bisheriger Fokus:

Wie frage ich Daten ab?

Heute:

Was frage ich ab, um Daten zu verstehen?

Explorative Datenanalyse

Systematisches Erkunden von Daten, bevor man Modelle baut oder Hypothesen testet.

Übergang:

Von SQL-Technik zu
Analyse-Denkweise

Definition (John Tukey, 1977):

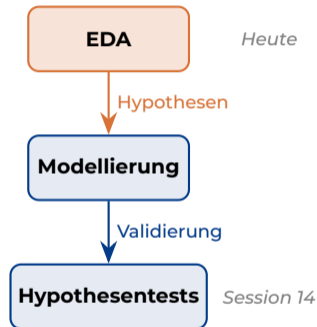
“Exploratory data analysis is detective work.”

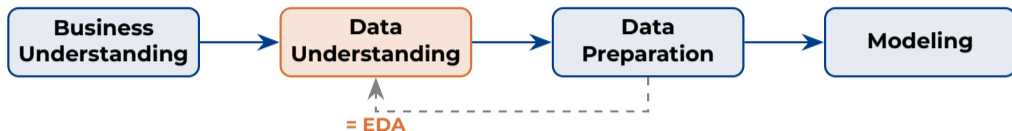
Ziele:

- 1 **Verstehen** – Was steckt in den Daten?
- 2 **Hypothesen generieren** – Welche Fragen ergeben sich?
- 3 **Qualität prüfen** – Fehler, Ausreißer, Missing Values?
- 4 **Vorbereiten** – Was braucht Bereinigung?

Wichtig:

EDA kommt vor konfirmatorischer Analyse!





EDA beantwortet:

- Wie viele Datensätze/Variablen?
- Welche Datentypen?
- Wie sind Werte verteilt?
- Gibt es Zusammenhänge?
- Was fehlt oder ist fehlerhaft?

Ohne EDA riskiert man:

- Falsche Modellauswahl
- Übersehene Datenprobleme
- Fehlinterpretationen
- Zeitverschwendung

Hands-on

Erste Dateninspektion

marimo: 11-eda.py

Aufgaben 11.1 – 11.2

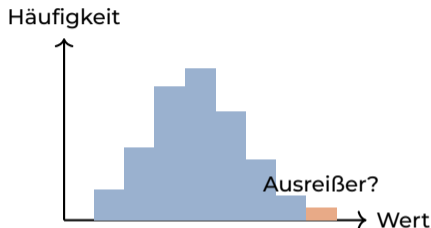
- 1 Motivation & EDA-Ziele
- **2 Univariate Analyse**
- 3 Bivariate Analyse
- 4 Visualisierung
- 5 SQL für EDA
- 6 Praktische EDA
- 7 Zusammenfassung

Definition:

Analyse *einer* Variable isoliert betrachtet.

Zentrale Fragen:

- Welche Werte kommen vor?
- Wie sind sie verteilt?
- Was ist "typisch"?
- Was ist ungewöhnlich?



Wichtige Kennzahlen:

Maß	Beschreibt	SQL
Mittelwert	Zentrum	AVG(x)
Median	Robustes Zentrum	PERCENTILE_CONT(0.5)
Min/Max	Wertebereich	MIN(x), MAX(x)
Standardabweichung	Streuung	STDDEV(x)

Mittelwert (Mean):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Berücksichtigt *alle* Werte
- Sensibel für Ausreißer
- Gut bei symmetrischen Verteilungen

Median:

- Mittlerer Wert (50% darüber/darunter)
- **Robust** gegen Ausreißer
- Gut bei schiefen Verteilungen

Beispiel: Gehälter

Mitarbeiter	Gehalt
A	45.000
B	48.000
C	52.000
D	55.000
E (CEO)	500.000
Mean	140.000
Median	52.000

Faustregel

Bei schiefen Verteilungen: Median \neq Mean

→ Immer beide berichten!

Spannweite (Range):

$$R = x_{max} - x_{min}$$

- Einfach zu berechnen
- Sehr sensibel für Ausreißer

Standardabweichung:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

- Mittlere Abweichung vom Mittelwert
- Gleiche Einheit wie Daten

Interquartilsabstand (IQR):

$$IQR = Q_3 - Q_1$$

- Bereich der mittleren 50%
- **Robust** gegen Ausreißer

Perzentile:

- Q_1 (25%): 25% der Werte darunter
- Q_2 (50%): Median
- Q_3 (75%): 75% der Werte darunter



Was ist ein Ausreißer?

Ein Wert, der “ungewöhnlich” weit vom Rest entfernt liegt.

IQR-Regel (Tukey):

Ausreißer, wenn:

$$x < Q_1 - 1.5 \cdot IQR$$

$$x > Q_3 + 1.5 \cdot IQR$$

Z-Score:

$$z = \frac{x - \bar{x}}{s}$$

Ausreißer, wenn $|z| > 3$

Aber Vorsicht!

- Ausreißer \neq Fehler
- Können wichtige Information sein
- Immer untersuchen, nie blind löschen

Beispiele für “echte” Ausreißer

- CEO-Gehalt in Gehaltsdaten
- Weihnachtsgeschäft in Umsatzdaten
- Seltene Krankheit in Patientendaten

Hands-on

Univariate Statistiken berechnen

marimo: 11-eda.py

Aufgaben 11.3 – 11.4

- 1 Motivation & EDA-Ziele
- 2 Univariate Analyse
- ▶ **3 Bivariate Analyse**
- 4 Visualisierung
- 5 SQL für EDA
- 6 Praktische EDA
- 7 Zusammenfassung

Definition:

Analyse der Beziehung zwischen zwei Variablen.

Zentrale Fragen:

- Hängen die Variablen zusammen?
- Wie stark ist der Zusammenhang?
- In welche Richtung?



Abhängig vom Variablentyp:

X	Y	Methode
Numerisch	Numerisch	Korrelation, Scatterplot
Kategorisch	Numerisch	Gruppenvergleich, Boxplots
Kategorisch	Kategorisch	Kreuztabelle, Chi-Quadrat

Pearson-Korrelationskoeffizient:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

Interpretation:

- $r = 1$: Perfekt positiv
- $r = 0$: Kein linearer Zusammenhang
- $r = -1$: Perfekt negativ

Faustregel:

$ r < 0.3$	schwach
$0.3 \leq r < 0.7$	mittel
$ r \geq 0.7$	stark

Wichtig!

Korrelation \neq Kausalität

Beispiel: Eisverkauf korreliert mit Ertrinkungsfällen.

→ Beides korreliert mit Temperatur!

In SQL:

```
SELECT CORR(x, y) AS korrelation  
FROM tabelle;
```

Fragestellung:

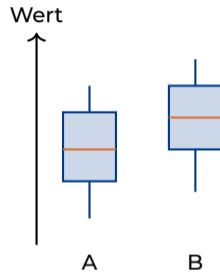
Unterscheiden sich Gruppen in einer numerischen Variable?

Beispiele:

- Gehalt nach Abteilung
- Punkte nach Position (Fußball)
- Umsatz nach Region

SQL-Ansatz:

```
SELECT gruppe ,  
       AVG(wert) AS mittel ,  
       STDDEV(wert) AS streuung  
FROM tabelle  
GROUP BY gruppe;
```



Boxplot:

Zeigt Median, IQR, Ausreißer
Ideal für Gruppenvergleiche

Hands-on

Zusammenhänge untersuchen

marimo: 11-eda.py

Aufgaben 11.5 – 11.6

Pause

15 Minuten

- 1 Motivation & EDA-Ziele
- 2 Univariate Analyse
- 3 Bivariate Analyse
- ▶ **4 Visualisierung**
- 5 SQL für EDA
- 6 Praktische EDA
- 7 Zusammenfassung

Histogramm
Verteilung

Boxplot
Verteilung +
Ausreißer

Liniendiagramm
Zeitreihen

Scatterplot
2 numerische
Variablen

Balkendiagramm
Kategorien
vergleichen

Heatmap
Korrelations-
matrix

Univariat:

- Histogramm: Verteilungsform
- Boxplot: Zusammenfassung + Ausreißer

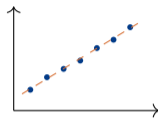
Bivariat/Multivariat:

- Scatterplot: Zusammenhang zweier Variablen
- Heatmap: Viele Korrelationen auf einmal

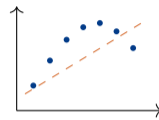
Vier Datensätze mit:

- Gleichem Mittelwert (X und Y)
- Gleicher Varianz
- Gleicher Korrelation ($r = 0.816$)
- Gleicher Regressionslinie

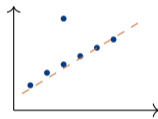
Aber völlig unterschiedlichen Mustern!



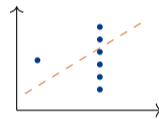
I: Linear



II: Kurve



III: Ausreißer



IV: Vertikal

Lektion

Immer visualisieren!

Statistiken allein können täuschen.

1. Datenüberblick

- ☐ Anzahl Zeilen/Spalten
- ☐ Datentypen prüfen
- ☐ Erste Zeilen ansehen

2. Missing Values

- ☐ Fehlende Werte zählen
- ☐ Muster erkennen (MCAR, MAR, MNAR?)
- ☐ Strategie festlegen

3. Univariate Statistiken

- ☐ Lagemaße (Mean, Median)
- ☐ Streuungsmaße (SD, IQR)
- ☐ Verteilungen visualisieren

4. Ausreißer

- ☐ Identifizieren (IQR, Z-Score)
- ☐ Untersuchen (Fehler vs. echt)
- ☐ Dokumentieren

5. Bivariate Analyse

- ☐ Korrelationen berechnen
- ☐ Scatterplots erstellen
- ☐ Gruppenvergleiche

6. Dokumentation

- ☐ Erkenntnisse festhalten
- ☐ Hypothesen formulieren
- ☐ Nächste Schritte planen

- 1 Motivation & EDA-Ziele
- 2 Univariate Analyse
- 3 Bivariate Analyse
- 4 Visualisierung
- ▶ **5 SQL für EDA**
- 6 Praktische EDA
- 7 Zusammenfassung

Grundlegende Statistiken:

```
SELECT
    COUNT(*) AS n,
    AVG(gehalt) AS mean,
    MIN(gehalt) AS min,
    MAX(gehalt) AS max,
    STDDEV(gehalt) AS std
FROM mitarbeiter;
```

Perzentile (DuckDB):

```
SELECT
    PERCENTILE_CONT(0.25)
        WITHIN GROUP (ORDER BY gehalt)
        AS q1,
    PERCENTILE_CONT(0.50)
        WITHIN GROUP (ORDER BY gehalt)
        AS median,
    PERCENTILE_CONT(0.75)
        WITHIN GROUP (ORDER BY gehalt)
```

Missing Values zählen:

```
SELECT
    COUNT(*) AS total,
    COUNT(gehalt) AS nicht_null,
    COUNT(*) - COUNT(gehalt)
        AS null_count
FROM mitarbeiter;
```

Korrelation:

```
SELECT
    CORR(alter, gehalt)
        AS korrelation
FROM mitarbeiter;
```

Binning (Werte gruppieren):

```
SELECT
    CASE
        WHEN gehalt < 40000
            THEN 'niedrig'
        WHEN gehalt < 70000
            THEN 'mittel'
        ELSE 'hoch'
    END AS gehaltsklasse,
    COUNT(*) AS anzahl
FROM mitarbeiter
GROUP BY gehaltsklasse;
```

Häufigkeitsverteilung:

```
SELECT
    abteilung,
    COUNT(*) AS n,
    ROUND(COUNT(*) * 100.0 /
        (SELECT COUNT(*)
         FROM mitarbeiter), 1)
        AS prozent
FROM mitarbeiter
GROUP BY abteilung
ORDER BY n DESC;
```

Tipp: NTILE für gleichmäßige Gruppen

NTILE(4) OVER (ORDER BY gehalt) teilt in 4 gleich große Quartile.

Mit IQR-Regel:

```
WITH quartile AS (  
    SELECT  
        PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY gehalt) AS q1,  
        PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY gehalt) AS q3  
    FROM mitarbeiter  
) ,  
grenzen AS (  
    SELECT  
        q1 - 1.5 * (q3 - q1) AS untere_grenze ,  
        q3 + 1.5 * (q3 - q1) AS obere_grenze  
    FROM quartile  
)  
SELECT m.*  
FROM mitarbeiter m, grenzen g  
WHERE m.gehalt < g.untere_grenze  
    OR m.gehalt > g.obere_grenze;
```

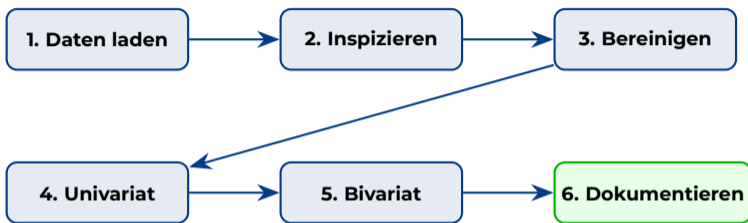
Hands-on

SQL für EDA anwenden

marimo: 11-eda.py

Aufgaben 11.7 – 11.8

- 1 Motivation & EDA-Ziele
- 2 Univariate Analyse
- 3 Bivariate Analyse
- 4 Visualisierung
- 5 SQL für EDA
- ▶ **6 Praktische EDA**
- 7 Zusammenfassung



Typische Erkenntnisse:

- Datenqualitätsprobleme
- Unerwartete Verteilungen
- Interessante Zusammenhänge
- Segmentierungsmöglichkeiten

Output einer EDA:

- Data Quality Report
- Deskriptive Statistiken
- Visualisierungen
- Hypothesen für weitere Analyse

Hands-on

Vollständige EDA durchführen

marimo: 11-eda.py

Aufgaben 11.9 – 11.12

Erweiterte Übungszeit: 40 Minuten

- 1 Motivation & EDA-Ziele
- 2 Univariate Analyse
- 3 Bivariate Analyse
- 4 Visualisierung
- 5 SQL für EDA
- 6 Praktische EDA
- ▶ **7 Zusammenfassung**

EDA-Grundlagen:

- Systematisches Datenverständnis
- Vor Modellierung/Hypothesentests
- Visualisierung essentiell

Univariate Analyse:

- Lagemaße: Mean, Median
- Streuung: SD, IQR
- Ausreißer: IQR-Regel

Bivariate Analyse:

- Korrelation (numerisch)
- Gruppenvergleiche (kategorisch)
- Korrelation \neq Kausalität!

SQL für EDA:

- Aggregatfunktionen
- CASE WHEN für Binning
- CTEs für Ausreißer-Erkennung

Anscombe's Quartet: **Immer visualisieren!**

Session 12: Zeitreihenanalyse

Themen:

- Zeitreihenkomponenten
- Trend, Saisonalität, Residuen
- Window Functions: LAG, LEAD
- Moving Averages
- Year-over-Year Vergleiche

Vorbereitung:

- Session 9 (JOINS) wiederholen
- OVER-Klausel aus CTEs

Vorschau

LAG(sales, 12) OVER (ORDER BY
month)
→ Wert vor 12 Monaten

Fragen?

christopher.flath@uni-wuerzburg.de