

Sommersemester 2026

Datenmanagement & -analyse

Prof. Dr. Christoph M. Flath

Data Driven Decisions Group, Universität Würzburg

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- 3 Fallstudie I: Dr. Harold Shipman
- 4 SQL-Analyse des Shipman-Falls
- 5 Fallstudie II: Benford's Law
- 6 Betrugserkennung mit Ziffernanalyse
- 7 Zusammenfassung & Ausblick

Lernziele

Strukturierte Datenanalyseprozesse verstehen und auf reale Fallstudien anwenden.

► 1 **Rückblick & CRISP-DM Einführung**

2 Der CRISP-DM Prozess

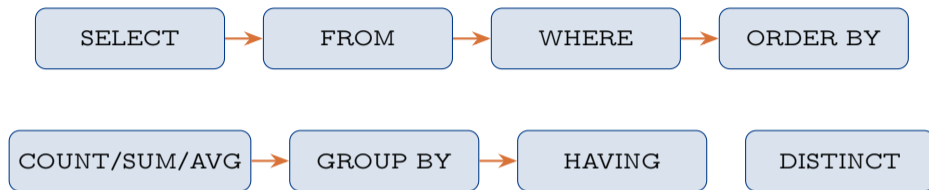
3 Fallstudie I: Dr. Harold Shipman

4 SQL-Analyse des Shipman-Falls

5 Fallstudie II: Benford's Law

6 Betrugserkennung mit Ziffernanalyse

7 Zusammenfassung & Ausblick



Bisher gelernt:

- Daten abfragen, filtern und sortieren
- Aggregieren und gruppieren
- NULL-Werte behandeln

Heute: Diese Werkzeuge systematisch einsetzen!

Bisher: Einzelne Abfragen

- “Zeige alle Spieler”
- “Wie viele Tore?”
- “Gruppiere nach Position”

⇒ Isolierte technische Übungen

Heute: Strukturierte Analyse

- Geschäftsproblem verstehen
- Daten systematisch erkunden
- Hypothesen prüfen
- Erkenntnisse kommunizieren

⇒ Professionelle Datenanalyse

Kernfrage

Wie gehen professionelle Datenanalysten systematisch vor?

Ohne Struktur:

- Ad-hoc Abfragen
- Wichtige Fragen übersehen
- Ergebnisse nicht reproduzierbar
- Fehlinterpretationen
- Zeit verschwendet

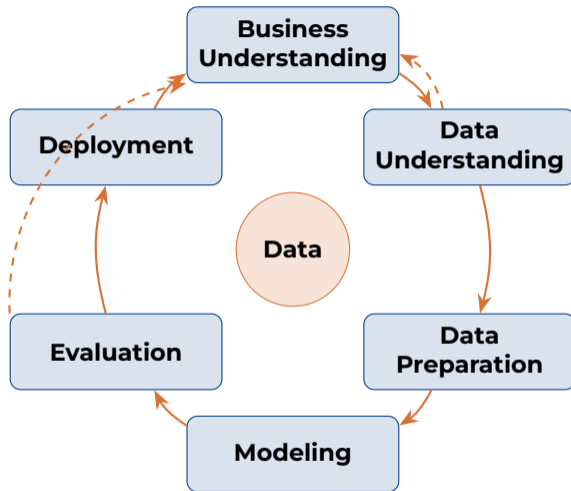
Mit Struktur:

- Klare Zielsetzung
- Systematische Exploration
- Dokumentierte Schritte
- Validierte Ergebnisse
- Effiziente Arbeit

Realität

80% der Analysezeit geht für Datenvorbereitung drauf –
ein guter Prozess hilft, diese Zeit sinnvoll zu nutzen.

- 1 Rückblick & CRISP-DM Einführung
- ▶ **2 Der CRISP-DM Prozess**
- 3 Fallstudie I: Dr. Harold Shipman
- 4 SQL-Analyse des Shipman-Falls
- 5 Fallstudie II: Benford's Law
- 6 Betrugserkennung mit Ziffernanalyse
- 7 Zusammenfassung & Ausblick



Industrie-Standard seit 1996 – entwickelt von IBM, NCR, SPSS, Daimler

Kernfragen:

- Was ist das Geschäftsziel?
- Welches Problem soll gelöst werden?
- Wie wird Erfolg gemessen?
- Wer sind die Stakeholder?

Output:

- Klare Problemdefinition
- Erfolgskriterien
- Projektplan

Beispiel: Shipman

“Gibt es Ärzte mit ungewöhnlich hohen Sterberaten?”

⇒ Anomalie-Erkennung

⇒ Patientensicherheit

Aktivitäten:

- Daten sammeln
- Daten beschreiben (Struktur, Umfang)
- Daten erkunden (Verteilungen, Muster)
- Datenqualität prüfen

SQL-Werkzeuge:

- `SELECT *` – Struktur verstehen
- `COUNT, DISTINCT` – Umfang
- `MIN, MAX, AVG` – Verteilungen
- `IS NULL` – Datenqualität

Häufige Probleme

- Fehlende Werte
- Inkonsistente Formate
- Duplikate
- Ausreißer

Die arbeitsintensivste Phase!

Typische Aufgaben:

- Daten bereinigen
- Fehlende Werte behandeln
- Variablen transformieren
- Daten zusammenführen
- Stichproben ziehen

SQL-Beispiele:

- COALESCE für NULL
- CAST für Typumwandlung
- SUBSTR für Textextraktion
- JOIN für Verknüpfungen
- WHERE für Filter

Faustregel

Plane 60-80% der Projektzeit für Datenvorbereitung ein!

Vollständigkeit
Keine fehlenden Werte

Korrektheit
Werte sind richtig

Konsistenz
Keine Widersprüche

Aktualität
Daten sind aktuell

Eindeutigkeit
Keine Duplikate

Gültigkeit
Format/Wertebereich OK

“Garbage In, Garbage Out”

Die beste Analyse ist wertlos, wenn die Eingabedaten schlecht sind.
Datenqualität ist keine Option, sondern Voraussetzung für jeden Analyseerfolg!

Modeling:

- Technik wählen
- Modell bauen
- Parameter optimieren

*(Oft mit ML-Tools,
aber auch SQL-Analysen)*

Evaluation:

- Ergebnisse prüfen
- Geschäftsziele erreicht?
- Nächste Schritte?

*(Zurück zu Phase 1
wenn nötig)*

Deployment:

- Lösung einsetzen
- Dokumentieren
- Monitoring planen

*(Dashboards, Reports,
automatisierte Abfragen)*

Wichtig

CRISP-DM ist iterativ – man springt oft zwischen Phasen!

Ordne die Aktivitäten den CRISP-DM Phasen zu:

Aktivität	Phase?
“Wir wollen Kundenabwanderung reduzieren”	???
NULL-Werte mit Durchschnitt ersetzen	???
COUNT und AVG über alle Spalten	???
Dashboard für Management erstellen	???
Modell mit Testdaten validieren	???

Ordne die Aktivitäten den CRISP-DM Phasen zu:

Aktivität	Phase?
“Wir wollen Kundenabwanderung reduzieren”	???
NULL-Werte mit Durchschnitt ersetzen	???
COUNT und AVG über alle Spalten	???
Dashboard für Management erstellen	???
Modell mit Testdaten validieren	???

Lösung: Business Understanding, Data Preparation, Data Understanding, Deployment, Evaluation

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- ▶ **3 Fallstudie I: Dr. Harold Shipman**
- 4 SQL-Analyse des Shipman-Falls
- 5 Fallstudie II: Benford's Law
- 6 Betrugserkennung mit Ziffernanalyse
- 7 Zusammenfassung & Ausblick

Hintergrund:

- Britischer Hausarzt (1946-2004)
- Praktizierte in Hyde, Greater Manchester
- Über 23 Jahre als Arzt tätig
- Beliebter und respektierter Mediziner

Die erschreckende Wahrheit:

- Mind. 215 Patienten ermordet
- Meist ältere Frauen
- Tötung durch Morphin-Überdosis
- Produktivster Serienmörder der UK-Geschichte

Kernfrage

Hätte Datenanalyse die Morde früher aufdecken können?

Spoiler

Ja – die Muster waren in den Daten sichtbar!

1998: Der letzte Mord

- Kathleen Grundy (81) stirbt
- Shipman fälscht ihr Testament
- Tochter (Anwältin!) wird misstrauisch
- Exhumierung ⇒ Morphin nachgewiesen

Untersuchung danach:

- Statistische Analyse aller Todesfälle
- Shipmans Raten **extrem** auffällig
- Muster über Jahrzehnte erkennbar

1975-1998

23 Jahre aktiv

215+ Opfer

Nachgewiesen

1998 entdeckt

Durch Gier, nicht Daten

Geschäftsproblem:

- Patientensicherheit gewährleisten
- Ungewöhnliche Sterbemuster erkennen
- Ärzte mit Anomalien identifizieren

Analysefragen:

- 1 Wie viele Patienten sterben pro Arzt?
- 2 Gibt es Ärzte mit überdurchschnittlich vielen Todesfällen?
- 3 Zu welchen Uhrzeiten sterben die Patienten?
- 4 Gibt es Muster bei Alter oder Geschlecht der Verstorbenen?

Erfolgskriterium

Identifikation von Ärzten, deren Sterberaten signifikant vom Durchschnitt abweichen.

Verfügbare Informationen (vereinfacht):

Tabelle	Spalten	Beschreibung
deaths	patient_id, doctor_id, death_date, death_time, age, gender, cause	Todesfälle
doctors	doctor_id, name, practice, start_year	Ärzte
patients	patient_id, doctor_id, registration_date	Patientenstamm

Frage: Welche SQL-Abfragen würden Sie stellen?

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- 3 Fallstudie I: Dr. Harold Shipman
- **4 SQL-Analyse des Shipman-Falls**
- 5 Fallstudie II: Benford's Law
- 6 Betrugserkennung mit Ziffernanalyse
- 7 Zusammenfassung & Ausblick

Data Understanding: Wie verteilen sich die Todesfälle?

```
SELECT
    d.name AS arzt ,
    COUNT(*) AS todesfaelle
FROM deaths de
JOIN doctors d ON de.doctor_id = d.doctor_id
GROUP BY d.doctor_id, d.name
ORDER BY todesfaelle DESC;
```

Erwartetes Ergebnis:

arzt	todesfaelle
Dr. Shipman	297
Dr. Smith	45
Dr. Jones	42
...	...

⇒ **Shipman: 6-7x mehr als Kollegen!**

Hypothese: Natürliche Todesfälle verteilen sich über den Tag

```
SELECT
  doctor_id ,
  CASE
    WHEN death_time BETWEEN '09:00' AND '17:00'
    THEN 'Praxiszeit'
    ELSE 'Ausserhalb'
  END AS zeitraum ,
  COUNT(*) AS anzahl
FROM deaths
GROUP BY doctor_id , zeitraum ;
```

Normale Ärzte:

Praxiszeit: 35%

Außerhalb: 65%

Shipman:

Praxiszeit: **83%**

Außerhalb: 17%

```
SELECT
    d.name AS arzt,
    gender AS geschlecht,
    COUNT(*) AS anzahl,
    ROUND(COUNT(*) * 100.0 / SUM(COUNT(*)) OVER
        (PARTITION BY d.name), 1) AS prozent
FROM deaths de
JOIN doctors d ON de.doctor_id = d.doctor_id
GROUP BY d.name, gender;
```

Normale Verteilung:

Männlich: $\approx 48\%$

Weiblich: $\approx 52\%$

Shipman:

Männlich: 19%

Weiblich: **81%**

⇒ Starke Abweichung bei Geschlechterverteilung!

```
SELECT
    d.name AS arzt ,
    AVG(de.age) AS durchschnittsalter ,
    MIN(de.age) AS juengstes_opfer ,
    MAX(de.age) AS aeltestes_opfer
FROM deaths de
JOIN doctors d ON de.doctor_id = d.doctor_id
GROUP BY d.name;
```

Auffälligkeit:

- Shipmans Patienten: Durchschnitt 77 Jahre
- Aber: Auch “gesunde” 65-Jährige
- Typisch für Mord: Opferprofil ist konsistent

Aufgabe: Schreibe eine SQL-Abfrage, die zeigt:

- 1 Wie viele Patienten pro Arzt in den **ersten 6 Monaten** nach Registrierung sterben
- 2 Gruppiert nach Arzt
- 3 Sortiert nach Anzahl (absteigend)

Hinweis: Neue Patienten sollten eigentlich nicht so schnell sterben...

Aufgabe: Schreibe eine SQL-Abfrage, die zeigt:

- 1 Wie viele Patienten pro Arzt in den **ersten 6 Monaten** nach Registrierung sterben
- 2 Gruppiert nach Arzt
- 3 Sortiert nach Anzahl (absteigend)

Hinweis: Neue Patienten sollten eigentlich nicht so schnell sterben...

```
SELECT d.name, COUNT(*) AS fruehe_tode
FROM deaths de
JOIN doctors d ON de.doctor_id = d.doctor_id
JOIN patients p ON de.patient_id = p.patient_id
WHERE de.death_date <= p.registration_date + 180
GROUP BY d.name
ORDER BY fruehe_tode DESC;
```

Was die Daten zeigten:

- Extreme Sterberate
- Ungewöhnliche Uhrzeiten
- Bias bei Geschlecht/Alter
- Muster über Jahre konsistent

Warum wurde es übersehen?

- Keine systematische Analyse
- Daten nicht verknüpft
- Kein Monitoring-System

Konsequenz

UK führte nach Shipman systematisches Mortality Monitoring ein.

SQL-Abfragen wie unsere laufen jetzt **automatisch!**

CRISP-DM Lektion

Data Understanding hätte das Problem aufgedeckt – wenn jemand hingeschaut hätte.

10 Minuten Pause

Nach der Pause:
Benford's Law – Wie Zahlen Betrug verraten

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- 3 Fallstudie I: Dr. Harold Shipman
- 4 SQL-Analyse des Shipman-Falls
- **5 Fallstudie II: Benford's Law**
- 6 Betrugserkennung mit Ziffernanalyse
- 7 Zusammenfassung & Ausblick



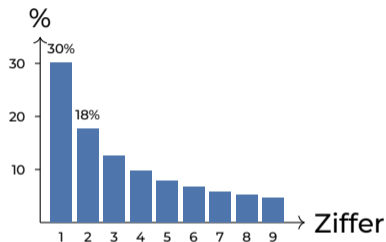
Die Beobachtung (1881):

Simon Newcomb bemerkte, dass Logarithmentafeln vorne abgegriffener waren als hinten.

Frank Benford (1938):

Analysierte 20.000+ Zahlen aus verschiedensten Quellen:

- Flusslängen
- Bevölkerungszahlen
- Physikalische Konstanten
- Zeitungsartikel



Erwartung: Jede Ziffer $\approx 11\%$

Realität: 1 kommt 6x häufiger vor als 9!

Mathematische Formel:

$$P(d) = \log_{10} \left(1 + \frac{1}{d} \right)$$

Erste Ziffer	1	2	3	4	5	6	7	8	9
Erwartete %	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

Warum funktioniert das?

- Natürliche Daten wachsen oft **multiplikativ** (exponentiell)
- Um von 1xx auf 2xx zu kommen: +100% Wachstum nötig
- Um von 8xx auf 9xx zu kommen: nur +12.5% Wachstum nötig
- \Rightarrow Zahlen "verbringen mehr Zeit" mit führender 1

Funktioniert gut bei:

- Finanzdaten (Rechnungen, Ausgaben)
- Bevölkerungszahlen
- Aktienkurse
- Naturwissenschaftliche Messungen
- Wahlergebnisse (Stimmen pro Bezirk)

⇒ Daten, die über mehrere Größenordnungen variieren

Funktioniert NICHT bei:

- Telefonnummern (feste Struktur)
- Postleitzahlen (zugewiesen)
- Preise (psychologisch: 9,99€)
- Körpergrößen (enger Bereich)
- Zufallszahlen

⇒ Daten mit eingeschränktem Wertebereich oder menschlichem Design

Die Idee:

Menschen, die Zahlen **erfinden**, verteilen Ziffern oft gleichmäßig – oder vermeiden “auffällige” Muster.

Echte Spesenabrechnungen:

- Folgen Benford
- Viele kleine Beträge
- Natürliche Variation

Gefälschte Abrechnungen:

- Zu viele 5er und 6er
- “Runde” Beträge
- Künstliche Gleichverteilung

Anwendungen

- Steuerbetrug (IRS nutzt Benford!)
- Bilanzbetrug (Enron wurde damit analysiert)
- Wissenschaftlicher Betrug (gefälschte Daten)
- Wahlbetrug

Welche Datensätze folgen Benford's Law?

Datensatz	Benford?
Umsätze aller DAX-Unternehmen	???
Hausnummern in Würzburg	???
Einwohnerzahlen aller Länder	???
Instagram-Follower von Influencern	???
Lottozahlen der letzten 10 Jahre	???

Welche Datensätze folgen Benford's Law?

Datensatz	Benford?
Umsätze aller DAX-Unternehmen	???
Hausnummern in Würzburg	???
Einwohnerzahlen aller Länder	???
Instagram-Follower von Influencern	???
Lottozahlen der letzten 10 Jahre	???

Lösungen: Ja, Nein (zugewiesen), Ja, Ja (wachsen exponentiell), Nein (Zufall im festen Bereich)

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- 3 Fallstudie I: Dr. Harold Shipman
- 4 SQL-Analyse des Shipman-Falls
- 5 Fallstudie II: Benford's Law
- **6 Betrugserkennung mit Ziffernanalyse**
- 7 Zusammenfassung & Ausblick

Schritt 1: Die erste Ziffer jeder Zahl ermitteln

```
-- Variante 1: String-Manipulation
SELECT
    betrag,
    CAST(SUBSTR(CAST(ABS(betrag) AS TEXT), 1, 1) AS INTEGER)
        AS erste_ziffer
FROM rechnungen
WHERE betrag > 0;
```

```
-- Variante 2: Mathematisch (eleganter)
SELECT
    betrag,
    CAST(ABS(betrag) / POWER(10, FLOOR(LOG10(ABS(betrag)))))
        AS INTEGER) AS erste_ziffer
FROM rechnungen
WHERE betrag > 0;
```

```
SELECT
    CAST(SUBSTR(CAST(betrag AS TEXT), 1, 1) AS INTEGER)
        AS erste_ziffer ,
    COUNT(*) AS anzahl,
    ROUND(COUNT(*) * 100.0 / (SELECT COUNT(*) FROM rechnungen
        WHERE betrag > 0), 1) AS prozent
FROM rechnungen
WHERE betrag > 0
GROUP BY erste_ziffer
ORDER BY erste_ziffer;
```

Beispiel-Output:

erste_ziffer	anzahl	prozent
1	3012	30.1
2	1761	17.6
...

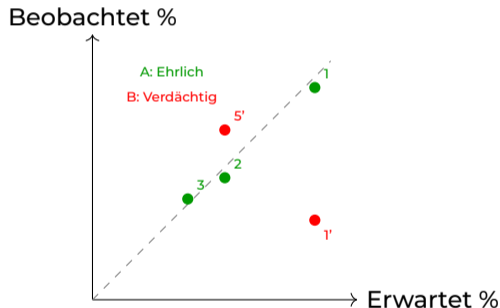
```
WITH ziffern AS (  
    SELECT CAST(SUBSTR(CAST(betrag AS TEXT), 1, 1) AS INT) AS d,  
           COUNT(*) AS n  
    FROM rechnungen WHERE betrag > 0  
    GROUP BY d  
) ,  
benford AS (  
    SELECT 1 AS d, 0.301 AS erwartet UNION ALL  
    SELECT 2, 0.176 UNION ALL  
    SELECT 3, 0.125 UNION ALL  
    -- ... weitere Ziffern  
)  
SELECT  
    z.d AS ziffer ,  
    z.n AS beobachtet ,  
    ROUND(b.erwartet * (SELECT SUM(n) FROM ziffern), 0) AS erwartet ,  
    ROUND(ABS(z.n - b.erwartet * (SELECT SUM(n) FROM ziffern)), 0)  
    AS abweichung  
FROM ziffern z  
JOIN benford b ON z.d = b.d;
```

Streudiagramm zeigt Abweichung von Benford auf einen Blick:

```
px.scatter(vergleich ,  
          x="erwartet_pct",  
          y="beobachtet_pct",  
          text="ziffer",  
          title="Benford - Check")
```

Interpretation:

- Punkte auf Diagonale = OK
- Abweichung = Verdacht
- Je weiter weg, desto auffälliger



Szenario:

Ein Mitarbeiter reicht monatlich Spesenabrechnungen ein. Die Buchhaltung wird misstrauisch.

Mitarbeiter A (ehrllich):

Ziffer	Anteil
1	29.5%
2	18.2%
3	11.8%
...	...

✓ Passt zu Benford

Mitarbeiter B (verdächtig):

Ziffer	Anteil
1	11.2%
2	10.8%
3	12.1%
...	...

× Fast gleichverteilt – Warnsignal!

Gegeben: Tabelle `invoices` mit Spalte `amount`

Aufgabe: Schreibe eine Abfrage, die:

- 1 Die erste Ziffer jedes Betrags extrahiert
- 2 Die Häufigkeit pro Ziffer zählt
- 3 Den Prozentanteil berechnet
- 4 Nach Ziffer sortiert

Gegeben: Tabelle invoices mit Spalte amount

Aufgabe: Schreibe eine Abfrage, die:

- 1 Die erste Ziffer jedes Betrags extrahiert
- 2 Die Häufigkeit pro Ziffer zählt
- 3 Den Prozentanteil berechnet
- 4 Nach Ziffer sortiert

```
SELECT
    CAST(SUBSTR(CAST(amount AS TEXT), 1, 1) AS INT) AS digit ,
    COUNT(*) AS count ,
    ROUND(COUNT(*) * 100.0 / SUM(COUNT(*)) OVER(), 1) AS pct
FROM invoices
WHERE amount > 0
GROUP BY digit
ORDER BY digit;
```

Enron (2001):

- Bilanzfälschung
- Finanzberichte wichen von Benford ab
- Bestimmte Ziffern unterrepräsentiert

Griechenland (2011):

- Wirtschaftsdaten manipuliert
- Benford-Analyse zeigte Anomalien
- EU-Beitritts-Statistiken verfälscht

Wahlen (diverse):

- Iran 2009
- Venezuela
- Russland

Benford-Abweichungen als Indikator für mögliche Manipulation.

Wichtig

Benford ist ein **Indikator**, kein Beweis! Abweichungen erfordern weitere Untersuchung.

Praktische Anwendung: Audit Analytics

Typische Prüfungsbereiche:

- Reisekostenabrechnungen
- Lieferantenrechnungen
- Lagerbestandsbewertungen
- Umsatzerlöse
- Kreditorenbuchhaltung

Warnsignale:

- Zu viele "runde" Beträge
- Häufung knapp unter Genehmigungsgrenze (z.B. viele 49,90€ statt 50€+)
- Gleichverteilung der Ziffern
- Fehlende kleine Beträge

Big Four nutzen Benford

Wirtschaftsprüfungsgesellschaften (Deloitte, PwC, EY, KPMG) setzen Benford-Analysen standardmäßig als Teil ihrer Audit-Prozeduren ein.

Phase	Anwendung
Business Understanding	Welche Daten könnten manipuliert sein? Spesen? Steuern? Wahlen?
Data Understanding	Sind die Daten für Benford geeignet? (Mehrere Größenordnungen?)
Data Preparation	Erste Ziffer extrahieren, Nullen/Negative behandeln
Modeling	Verteilung berechnen, mit Benford vergleichen
Evaluation	Statistische Signifikanz prüfen (Chi-Quadrat-Test)
Deployment	Automatisiertes Monitoring einrichten

Szenario: Du analysierst Todesfälle UND Rechnungsdaten einer Praxis.

Aufgabe 1: Welche Abfrage zeigt die Todesfälle pro Wochentag?

Szenario: Du analysierst Todesfälle UND Rechnungsdaten einer Praxis.

Aufgabe 1: Welche Abfrage zeigt die Todesfälle pro Wochentag?

```
SELECT
    strftime('%w', death_date) AS wochentag,
    COUNT(*) AS anzahl
FROM deaths
GROUP BY wochentag
ORDER BY wochentag;
```

Aufgabe 2: Was wäre auffällig, wenn die meisten Todesfälle an Montagen passieren?

Szenario: Du analysierst Todesfälle UND Rechnungsdaten einer Praxis.

Aufgabe 1: Welche Abfrage zeigt die Todesfälle pro Wochentag?

```
SELECT
    strftime('%w', death_date) AS wochentag,
    COUNT(*) AS anzahl
FROM deaths
GROUP BY wochentag
ORDER BY wochentag;
```

Aufgabe 2: Was wäre auffällig, wenn die meisten Todesfälle an Montagen passieren?

⇒ Praxisöffnung! Natürliche Todesfälle verteilen sich gleichmäßiger.

Diese Abfrage soll die Benford-Verteilung berechnen. Was ist falsch?

```
SELECT
    SUBSTR(betrag, 1, 1) AS erste_ziffer ,
    COUNT(*) AS anzahl
FROM rechnungen
GROUP BY erste_ziffer;
```

Diese Abfrage soll die Benford-Verteilung berechnen. Was ist falsch?

```
SELECT
    SUBSTR(betrag, 1, 1) AS erste_ziffer ,
    COUNT(*) AS anzahl
FROM rechnungen
GROUP BY erste_ziffer;
```

Probleme:

- 1 betrag ist eine Zahl – muss zu TEXT konvertiert werden
- 2 Negative Zahlen haben “-” als erste Ziffer
- 3 Nullen und kleine Dezimalzahlen (0.5) werden falsch behandelt

```
SELECT
    CAST(SUBSTR(CAST(ABS(betrag) AS TEXT), 1, 1) AS INT)
        AS erste_ziffer ,
    COUNT(*) AS anzahl,
    ROUND(COUNT(*) * 100.0 / SUM(COUNT(*)) OVER(), 1) AS prozent
FROM rechnungen
WHERE betrag > 0    -- Nur positive Beträge
    AND betrag >= 1 -- Keine Dezimalzahlen < 1
GROUP BY erste_ziffer
ORDER BY erste_ziffer;
```

Änderungen:

- ABS() für negative Werte
- CAST(...AS TEXT) für String-Extraktion
- Filter für positive Zahlen ≥ 1
- Prozentberechnung hinzugefügt

Datenanalyse kann:

- Leben retten (Shipman-Monitoring)
- Betrug aufdecken (Benford)
- Gerechtigkeit fördern
- Effizienz steigern

Aber auch:

- Unschuldige verdächtigen
- Falsche Schlüsse ziehen
- Bias verstärken
- Privatsphäre verletzen

Wichtige Fragen

- ① Sind meine Daten repräsentativ?
- ② Welche Fehler kann ich machen?
- ③ Wer trägt die Konsequenzen?
- ④ Ist Korrelation = Kausalität?

Merke

Anomalie \neq Schuld
Statistische Signifikanz \neq Praktische Relevanz

Welche Benford-Abweichung würdest du bei diesen Datensätzen erwarten?

- ① Unternehmensausgaben eines Start-ups (erste 2 Jahre)
- ② Einwohnerzahlen aller deutschen Gemeinden
- ③ Preise in einem Online-Shop (viele 9,99€, 19,99€...)
- ④ Bitcoin-Transaktionsbeträge

Welche Benford-Abweichung würdest du bei diesen Datensätzen erwarten?

- 1 Unternehmensausgaben eines Start-ups (erste 2 Jahre)
- 2 Einwohnerzahlen aller deutschen Gemeinden
- 3 Preise in einem Online-Shop (viele 9,99€, 19,99€...)
- 4 Bitcoin-Transaktionsbeträge

Erwartungen:

- 1 Eventuell Abweichungen (schnelles, ungleichmäßiges Wachstum)
- 2 Sollte Benford folgen (natürliche Verteilung)
- 3 Starke Abweichung (psychologische Preisgestaltung)
- 4 Sollte Benford folgen (exponentielles Wachstum, keine Manipulation)

- 1 Rückblick & CRISP-DM Einführung
- 2 Der CRISP-DM Prozess
- 3 Fallstudie I: Dr. Harold Shipman
- 4 SQL-Analyse des Shipman-Falls
- 5 Fallstudie II: Benford's Law
- 6 Betrugserkennung mit Ziffernanalyse
- **7 Zusammenfassung & Ausblick**

CRISP-DM:

- Strukturierter Analyseprozess
- 6 Phasen, iterativ
- Business Understanding zuerst!
- Data Preparation braucht Zeit

Shipman-Fall:

- Anomalien in Aggregaten
- Zeitliche Muster
- Demografische Abweichungen
- Monitoring kann Leben retten

Benford's Law:

- Erste-Ziffer-Verteilung
- 30% beginnen mit 1
- Gilt für natürliche Daten
- Abweichungen deuten auf Manipulation

SQL-Techniken:

- GROUP BY für Verteilungen
- CASE für Kategorisierung
- String-Funktionen (SUBSTR)
- Prozentberechnungen

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität			Phase
Chi-Quadrat-Test	auf	Benford-	???
Verteilung			

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität			Phase
Chi-Quadrat-Test auf Benford-Verteilung			???
			Modeling
"Wir wollen Betrugsfälle um 20% reduzieren"			???

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität			Phase
Chi-Quadrat-Test auf Benford-Verteilung			???
			Modeling
"Wir wollen Betrugsfälle um 20% reduzieren"			???
			Business Und.
Dashboard für Compliance-Abteilung bauen			???

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität	Phase
Chi-Quadrat-Test auf Benford-Verteilung	???
	Modeling
"Wir wollen Betrugsfälle um 20% reduzieren"	???
	Business Und.
Dashboard für Compliance-Abteilung bauen	???
	Deployment
Negative Beträge herausfiltern	???

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität	Phase
Chi-Quadrat-Test auf Benford-Verteilung	???
	Modeling
“Wir wollen Betrugsfälle um 20% reduzieren”	???
	Business Und.
Dashboard für Compliance-Abteilung bauen	???
	Deployment
Negative Beträge herausfiltern	???
	Data Preparation
Histogramm der Ziffernverteilung erstellen	???

Ordne jede Aktivität der richtigen CRISP-DM Phase zu:

Aktivität	Phase
Chi-Quadrat-Test auf Benford-Verteilung	???
	Modeling
“Wir wollen Betrugsfälle um 20% reduzieren”	???
	Business Und.
Dashboard für Compliance-Abteilung bauen	???
	Deployment
Negative Beträge herausfiltern	???
	Data Preparation
Histogramm der Ziffernverteilung erstellen	???

1. Welche CRISP-DM Phase kommt VOR dem Modellieren?

- A) Evaluation
- B) Data Preparation
- C) Deployment

1. Welche CRISP-DM Phase kommt VOR dem Modellieren?

- A) Evaluation
- B) Data Preparation
- C) Deployment

Antwort: B

2. Laut Benford's Law – welche erste Ziffer ist am häufigsten?

- A) 5 (die Mitte)
- B) 1 (etwa 30%)
- C) 9 (die größte)

1. Welche CRISP-DM Phase kommt VOR dem Modellieren?

- A) Evaluation
- B) Data Preparation
- C) Deployment

Antwort: B

2. Laut Benford's Law – welche erste Ziffer ist am häufigsten?

- A) 5 (die Mitte)
- B) 1 (etwa 30%)
- C) 9 (die größte)

Antwort: B

Vorlesung 5: Daten verknüpfen mit JOIN

- Mehrere Tabellen verbinden
- INNER JOIN, LEFT JOIN, RIGHT JOIN
- Komplexe Analysen über Tabellen hinweg
- Fallstudie: Lieferketten-Analyse

Vorgeschmack

Wie kombiniert man Kundendaten mit Bestellungen?
Wie findet man Kunden, die noch NIE bestellt haben?

Aufgabe für die Übungsphase:

- 1 Öffne das marimo-Notebook zu Vorlesung 4
- 2 Führe die Shipman-Analysen mit Beispieldaten durch
- 3 Berechne die Benford-Verteilung für einen Datensatz
- 4 Vergleiche die Ergebnisse mit der erwarteten Verteilung

Reflexionsfragen

- Welche anderen Anomalien hätte man bei Shipman finden können?
- In welchen Bereichen könnte eure zukünftige Firma Benford nutzen?
- Welche Grenzen hat die Ziffernanalyse?

Vor dem Start:

- Ziel klar definieren
- Stakeholder einbeziehen
- Erfolgskriterien festlegen
- Zeitplan realistisch planen (80% für Daten!)

Während der Analyse:

- Jeden Schritt dokumentieren
- Annahmen explizit machen
- Zwischenergebnisse prüfen
- Regelmäßig Feedback einholen

Bei der Präsentation:

- Geschäftsproblem zuerst
- Visualisierungen nutzen
- Limitationen benennen
- Handlungsempfehlungen geben

Goldene Regel

Korrelation \neq Kausalität

Nur weil zwei Dinge zusammenhängen, verursacht das eine nicht das andere!

Fragen?

Nächste Woche: JOINS – Daten verknüpfen

“Data is the new oil – but it needs to be refined.”