LU02a - Theorie

Ergänzende Theorie zur Datenerfassung, Datenanalyse und Datenbereinigung

1. Analyse der Daten (vor dem Bereinigen)

Bevor Daten bereinigt oder verarbeitet werden, müssen sie analysiert werden. Dabei geht es darum zu verstehen, welche Daten vorliegen und welche Probleme auftreten könnten. Typische Fragen bei der Datenanalyse sind:

- Welche Datenfelder (Spalten) gibt es?
- Welche Datentypen liegen vor? (z. B. Zahl, Text, Datum)
- Welche Werte kommen häufig vor? Gibt es Ausreisser?
- Gibt es fehlende oder offensichtlich falsche Werte?
- Sind Dubletten vorhanden (gleiche Datensätze mehrfach)?

Diese Voranalyse hilft zu entscheiden, welche Bereinigungen notwendig sind.

2. Datenformate

Damit Daten ausgewertet werden können, müssen sie im richtigen Format vorliegen. Beispiele:

- Zahlen: ohne Leerzeichen oder Buchstaben (z. B. „23.5“ statt „23,5kg“)
- Datumswerte: in einem einheitlichen Format (z. B. „2025-03-14“)
- Texte: einheitliche Schreibweisen (z. B. “männlich”, nicht “maennlich”)
- Ja/Nein-Daten: konsistent codiert (z. B. nur “1/0” oder nur “Ja/Nein”, nicht gemischt)

Je besser das Format, desto leichter kann eine Software die Daten korrekt interpretieren.

3. Typische Probleme bei der Datenerfassung

Bei der Erfassung – egal ob manuell oder elektronisch – können Fehler entstehen: Manuelle Erfassung

- Tippfehler
- unterschiedliche Schreibweisen
- falsche Zuordnung (z. B. Zahl in falsche Spalte)
- Lesefehler aus Originaldokumenten

Elektronische Erfassung

- Messfehler (z. B. Sensorfehler)
- technische Ausfälle / unvollständige Messungen
- falsch konfigurierte Geräte

Datenübernahme aus bestehenden Systemen

- Formatunterschiede (z. B. Anzahl als Text gespeichert)
- veraltete Daten
- Dubletten aus früheren Importen

4. Datenbereinigung

Die Datenbereinigung (engl. Data Cleaning) dient dazu, Daten korrekt, vollständig und einheitlich zu machen. Wichtige Schritte sind: a) Fehlerhafte Werte korrigieren

- Erkennbar falsche Werte anpassen (z. B. Temperatur „350°C“ statt „35.0°C“ → vermutlich Tippfehler)

b) Fehlende Werte behandeln

- Nachtragen, wenn Information bekannt ist
- Schätzen, falls sinnvoll (z. B. Durchschnitt verwenden)
- Löschen, wenn der Datensatz nicht weiterverwendbar ist

c) Dubletten entfernen

- doppelte Einträge identifizieren und zusammenführen oder löschen

d) Vereinheitlichung

- Schreibweisen angleichen (z. B. “ja”, “Ja”, “JA” → einheitlich “Ja”)
- einheitliche Formatierung (Datum, Kommazahlen, Masseinheiten)

e) Plausibilitätskontrolle Dies bedeutet, zu prüfen, ob die Werte realistisch sind:

- Alter von Personen zwischen 0 und 120 Jahren?
- Endzeit später als Startzeit?
- Messwerte innerhalb erwarteter Grenzen?

5. Ziel der Datenbereinigung

Nach der Bereinigung sollen die Daten …

- korrekt

- vollständig
- einheitlich
- logisch plausibel
- auswertbar

... sein, damit die anschliessende Analyse, z. B. mit Excel, Datenbanken oder Statistikprogrammen, zuverlässig funktioniert.

M162-LU01