ScaDS.AI DRESDEN LEIPZIG

CENTER FOR SCALABLE DATA ANALYTICS AND
ARTIFICIAL INTELLIGENCE

IOM AI School

TOPIC: Data Wrangling
SPEAKER: Matthias Täschner

GEFÖRDERT VOM
Bundesministerium
für Forschung, Technologie
und Raumfahrt

SACHSEN Diese Maßnahme wird gefördert durch die Bundesregierung aufgrund eines Beschlusses des Deutschen Bundestages. Diese Maßnahme wird mitfinanziert durch Steuermittel auf der Grundlage des von den Abgeordneten des Sächsischen Landtags beschlossenen Haushaltes.

Using materials from Laura Zigutyte, Michaela Unger (EKFZ, TU Dresden), Robert Haase (ScaDS.AI, Leipzig University)

What and Why?

- The process of cleaning, transforming, and structuring raw data into a usable format
- Real-world data is messy (missing values, outliers, duplicates, wrong formats)
- Better: clean, validated, structured data in compatible / common formats
- Preprocessing is essential before analysis, visualization, or machine learning

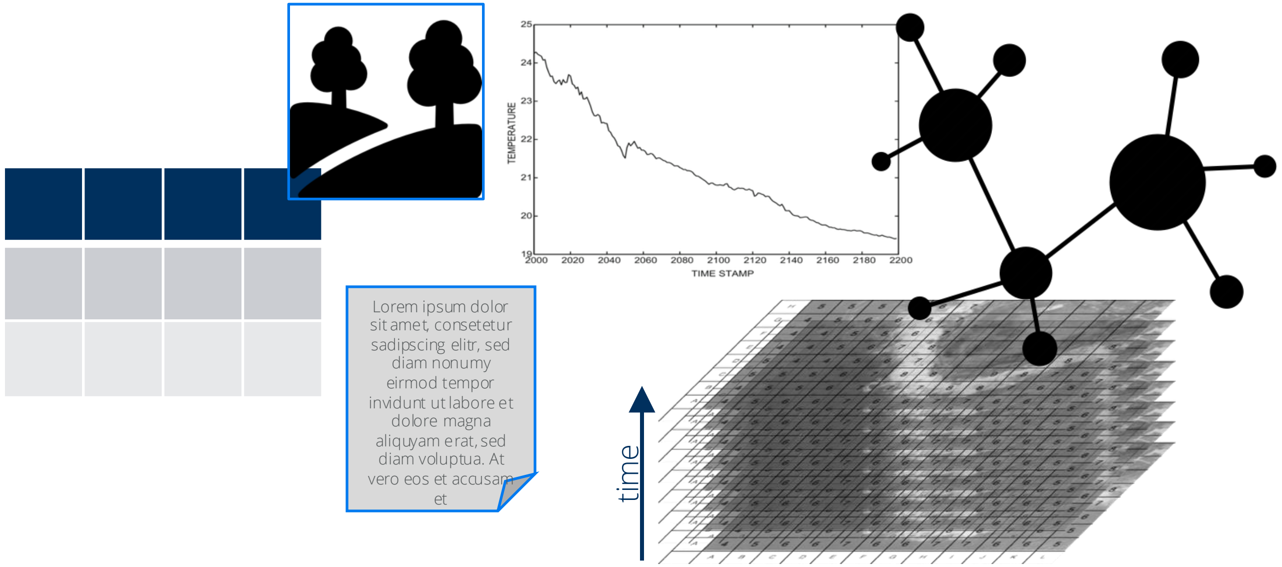
AGENDA

- Data Understanding
- Data Cleaning – Missing Data and Outliers
- Data Transformation
- Practice: Tabular data with pandas

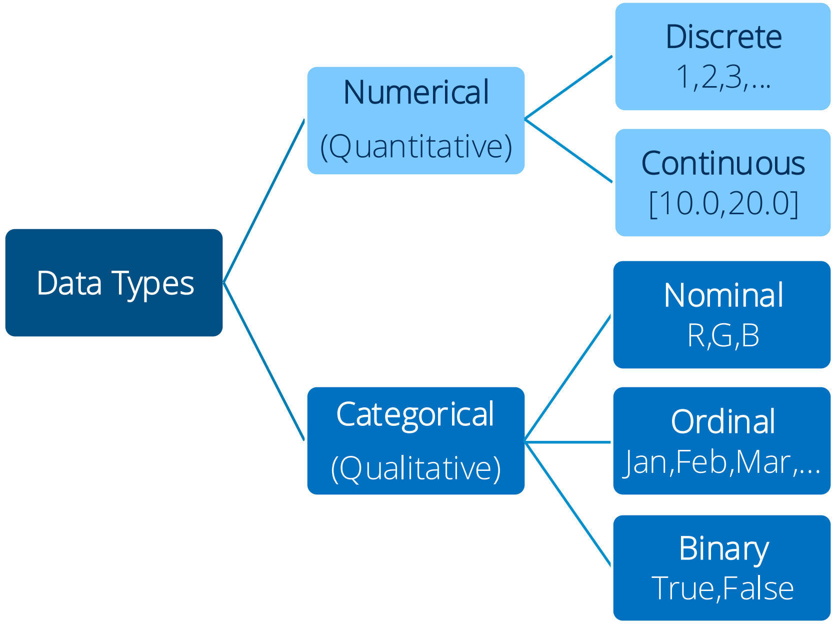
Data Understanding
Data Types and Structures

Know your data

- Understanding of the characteristics, differences, and specific requirements of different data types and structures
- Basis for selecting suitable data representations, processing and analysis procedures, and visualization methods



Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et



Data Types

Numerical (Quantitative)

Discrete 1,2,3,...

Continuous [10.0,20.0]

Categorical (Qualitative)

Nominal R,G,B

Ordinal Jan,Feb,Mar,...

Binary True,False

Source: own images

Data Cleaning
Missing Values

Definition

- Missing values in one or more variables/attributes of a data set
- Due to data entry errors, incomplete databases, data corruption, etc
- Potential systematic relationship between the probability for missing values in an attribute and the values of other attributes in the data set

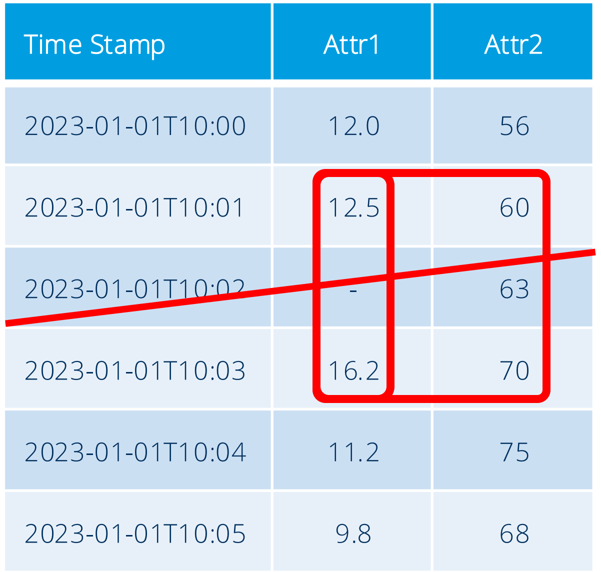
Effects

- Can lead to biased or inaccurate analysis results
- Analysis methods or ML models cannot handle missing values

Data Cleaning
Missing Values

Dealing with missing values

- Ignore
  - Taking missing values into account
  - Selecting suitable algorithms
- Delete
  - Removing data points with missing values
  - Reduced, but complete data set
- Interpolate
  - Estimating missing values based on existing values from the same attribute
- Impute
  - Estimating missing values using various techniques
  - Also based on the values of other attributes

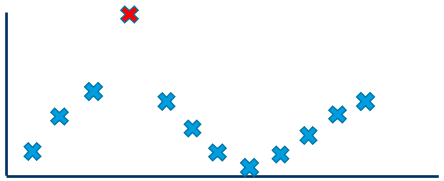


Time Stamp	Attr1	Attr2
2023-01-01T10:00	12.0	56
2023-01-01T10:01	12.5	60
~~2023-01-01T10:02~~	~~-~~	~~63~~
2023-01-01T10:03	16.2	70
2023-01-01T10:04	11.2	75
2023-01-01T10:05	9.8	68

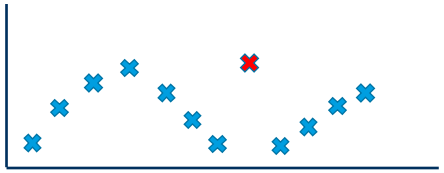
Data Cleaning
Outliers

Definition

- Values that deviate significantly from the majority of values in the data set
- Due to measurement, experimentation, or data entry errors, natural fluctuations in the data, or actual unusual observations
- Global outlier - compared to the entire data set



- Local outlier - compared to a subset



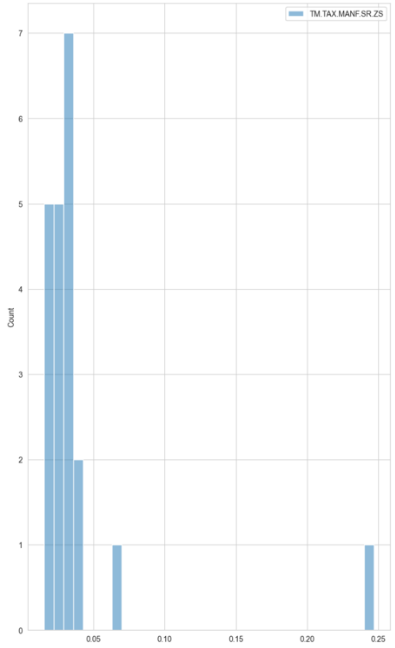
Effects

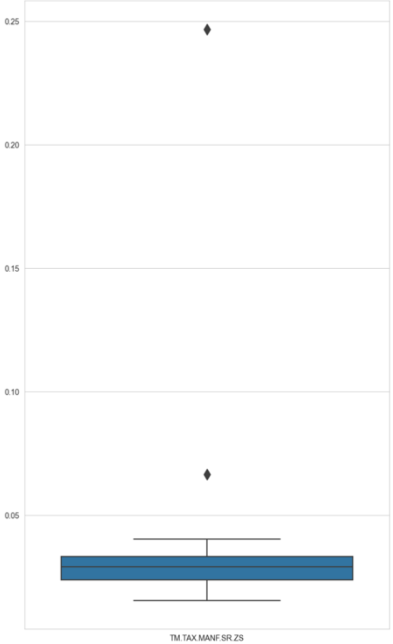
- Can skew statistical analyses, e.g., mean or standard deviation
- Can affect the performance of ML models, can lead to disturbed training, overfitting and poor generalizability, as well as reduced accuracy

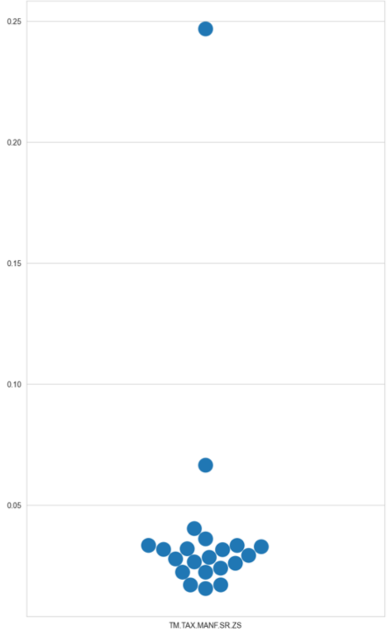
Data Cleaning
Outliers

How to detect outliers?

- Rule-based with thresholds or domain knowledge
- Statistical approaches such as Z-score
- ML-based approaches such as clustering
- Visual approaches such as histograms, box plots, or swarm plots







Source: own images

Data Cleaning
Outliers

Dealing with outliers

- Delete
  - Remove data points with outliers
- Capping/flooring
  - Replace with permissible predefined minimum or maximum
- Replace
  - Replace using methods such as interpolation, mean/median/most frequent value, regression
- Transform the data to mitigate the effect of outliers
  - Logarithmic transformation to reduce the magnitude/distortion in the data
  - Winsorization to limit outliers to a defined percentile of the data

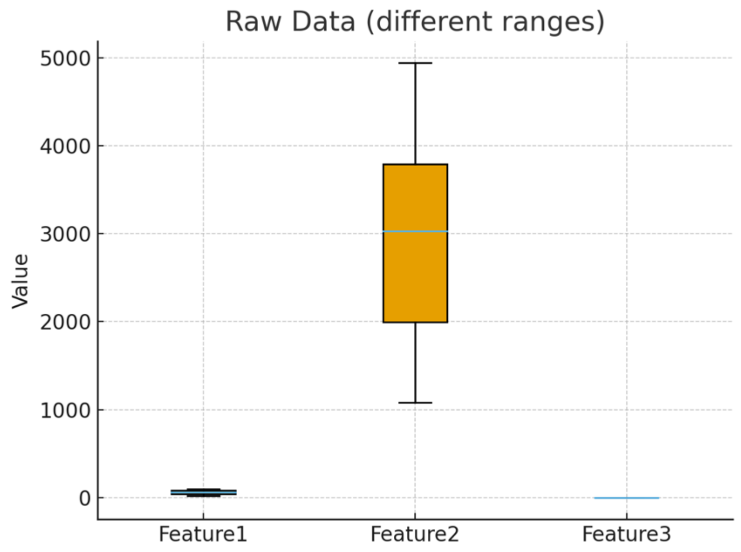
Data Transformation

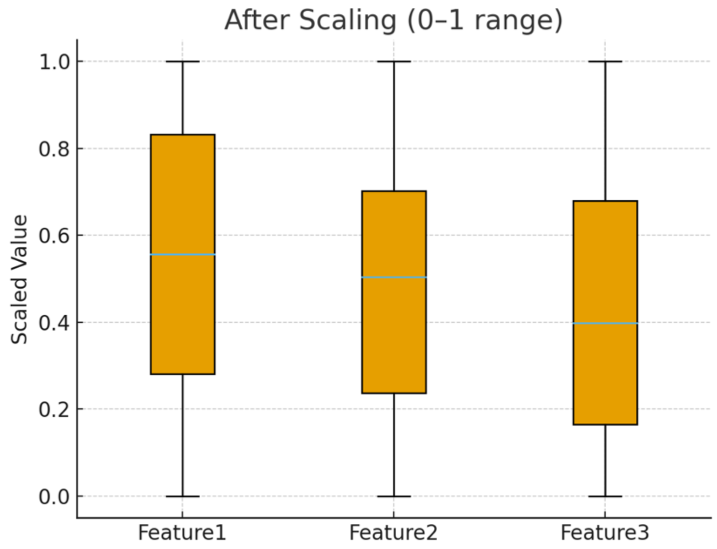
Why transform data?

- Make values comparable
- Improve performance of algorithms / ML models
- Reveal hidden patterns

Common transformations

- Scaling: rescale values to a fixed range, e.g., [0.0-1.0]
- Normalization: adjust distribution, e.g., z-score
- Numerical encoding of categories: one-hot, label encoding, ...
- Embeddings: numerical vector representation for text or images





Source: own images

What and Why? - Conclusion

- *The process of cleaning, transforming, and structuring raw data into a usable format*
- *Real-world data is messy (missing values, outliers, duplicates, wrong formats)*
- *Better: clean, validated, structured data in compatible / common formats*
- *Preprocessing is essential before analysis, visualization, or machine learning*

At the very least, consider all aspects mentioned and decide on an approach that is based on:

- Your data
- Your research question
- Your analysis goal and desired result or quality
- Your analysis method and algorithms
- …