BIG DATA & BUSINESS INTELLIGENCE

WEEK 3

Winter Semester 2025-2026

Lecturer: Narges Chinichian

SRH University of Applied Science

Feature Engineering for Data Analytics and Dashboards

Cleaning, Transforming, and Enriching Data Before Visualization

Foundation

What Is Feature Engineering?

- Transforming raw, imperfect data into meaningful, consistent information.
- In BI, it means preparing tables for reliable KPIs, visualizations, and aggregations.
- Typical steps: cleaning → standardizing → deriving → validating.

Why It Matters in BI

Garbage in = garbage out

Dashboards reflect your data quality.

Clean features → correct insights

Better decisions follow from reliable data.

Trust through consistency

Stakeholders trust your reports only if they see stable, reproducible numbers.

Analysis Journey

Key Message: Feature engineering is the bridge between ingestion and analysis.

Typical Data Quality Issues

- Missing or null values
- Inconsistent formats
  ("DE", "Germany", "ger")
- Duplicates and mixed types
- Outliers and unrealistic values
- Timezone and date inconsistencies

Step 1

Data Profiling/ Through EDA

Recap from last session:

Understand your dataset before touching it.

Use .info(), .describe(), .isna().sum(), .nunique().

Identify:

- Variable types
- % missing values
- Value ranges
- Potential duplicates

Step 2

Handling Missing Data

Decide

Drop, impute, or flag.

Numeric

Mean/median replacement.

Categorical

Mode or placeholder ("Unknown").

Never impute blindly, your imputation strategy matters, we will have a look at the most frequent imputation method in Notebook 2.

Step 3

Fixing Inconsistent Categories

01

Harmonize text entries

Strip spaces, unify case.

Berlin = BERLIN = berlin

02

Use mapping dictionaries

For country or product names.

GER, DE, de, Deutschland => Germany

03

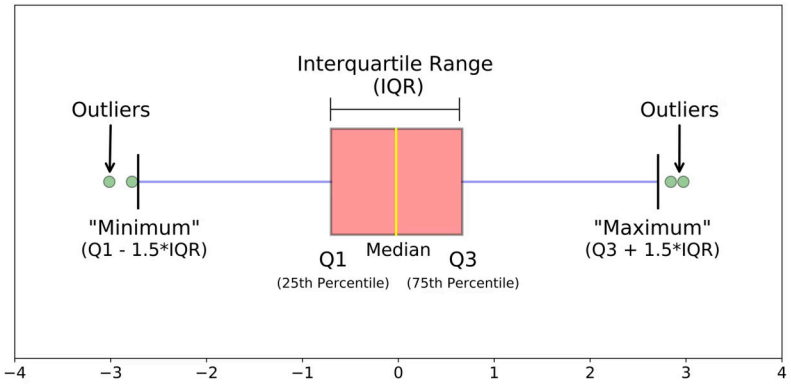
Validate category counts

Before/after cleaning.

Step 4

Dealing with Outliers

- Detect unusual numeric values using IQR or z-score.
- Decide whether to cap, replace, or ignore.
- Consider business meaning — is a 1 M € sale an error or a big client?



Step 5

Normalization & Scaling

Comparable ranges

Bring metrics to comparable ranges (e.g., revenue per employee).

Log-transform

Log-transform skewed distributions.

Preserve originals

Keep original values separately if they have interpretive value.

Step 6

Dates and Time Features

Convert to datetime types

Extract useful parts

Year, month, weekday, hour.

Flag special periods

Weekends, holidays, or fiscal periods for dashboard filters.

Step 7

Text Cleaning

- Lowercase, remove extra spaces, standardize units.
- Split compound fields ("City – Region").
- Extract keywords or domains if useful for grouping.

Step 8

Derived Metrics — Compute or Store?

Sometimes it's faster to compute *on demand* (profit = revenue – cost).

Only store if:

The transformation is heavy.

It's reused constantly.

The BI tool can't calculate it live.

Step 9

Aggregations and Groupings

- Aggregate where detail isn't needed (e.g., monthly totals).
- Use groupby in pandas.
- Define metrics once for consistency across dashboards.

Step 10

Validation Checks

1

No nulls in key fields

2

Unique identifiers confirmed

3

Correct data types

For each column.

4

Totals and subtotals match

Business expectations.

Documentation & Data Dictionary

Keep a log

Keep a log of each transformation.

Define meaning

Define meaning and unit of every column.

Enable reproducibility

Enables reproducibility and easier onboarding.

Hands-on Notebooks

1- Data Profiling & Cleaning Strategy

2- Imputation & Standardization

3- Data Integrity & Business Rules

4- Metric Design

5- Dashboard Dataset Preparation

Homework

Apply feature engineering on your datasets

Prepare and justify your KPIs

Present your results

Next session: We start with your presentation and wireframing.