4. Data Processing: Cleaning, EDA, and Feature Engineering

Cleaning

- Remove useless features.
- Null values.
    - Drop whole rows.
    - Drop whole columns.
    - Impute: mean, median, mode, ML.
- Duplicated values.
    - Are they erroneous? → Remove them.
    - Are they valid? → Decide what to do.
- Outliers.
    - Are they erroneous? → Remove or impute them.
    - Are they valid? → Decide what to do.

Exploratory Data Analysis (EDA)

- Column distributions.
- Correlations between features.
- Relations between features (e.g., pair-plot).
- Correlations with the target feature.
- Any interesting insights about the data.

Feature Engineering

- Encode categorical features.
    - One-hot encoding.
    - Ordinal encoding.
    - Others?
- Binning?
- Remove highly correlated features?
- Create new features that can be useful for prediction.