Data Preprocessing Techniques

- Handling missing values (removal, imputation: mean/median/mode, interpolation)
- Handling duplicate data
- Handling outliers (detection, removal, transformation)
- Encoding categorical variables
    - Label encoding
    - One-hot encoding
    - Ordinal encoding
    - Binary encoding
- Converting text to features (text vectorization: Bag-of-Words, TF-IDF, embedding basics)
- Date/time feature engineering (extracting day, month, weekday, etc.)
- Feature extraction and construction (creating new features, polynomial features, interaction terms)
- Dimensionality reduction (PCA, t-SNE, UMAP, feature selection)
- Data type conversion (string to numeric, etc.)

Feature Scaling & Normalization

- Min-Max scaling (normalization to [0, 1] range)
- Standardization (z-score normalization)
- Robust scaling (using median and IQR)
- Log transformation, Box-Cox, Yeo-Johnson transformations
- Unit vector normalization (L2 norm, L1 norm)
- When/why to use each scaling method (for which algorithms, e.g., SVM, KNN)

Exploratory Data Analysis (EDA)

- Descriptive statistics (mean, median, mode, std, skewness, kurtosis)
- Data visualization
    - Histograms
    - Boxplots and violin plots
    - Scatter plots
    - Pairplots (seaborn)
    - Heatmaps (correlation matrix)
    - Count plots, bar plots, pie charts (for categories)
- Distribution analysis (normality, skewness, kurtosis)
- Correlation analysis (Pearson, Spearman, Kendall)

- Groupby and aggregation
- Outlier detection (visual and statistical)
- Missing data visualization (missingno, heatmaps)
- Target variable analysis (distribution, imbalance)
- Feature importance analysis (tree-based, permutation, SHAP/ELI5 basics)

Other Related Techniques

- Data splitting (train-test split, stratified sampling, cross-validation)
- Data balancing (over/under-sampling, SMOTE)
- Pipeline creation (using sklearn's Pipeline)
- Data leakage detection and prevention
- Data augmentation (for images, text, etc.)
- Saving/loading preprocessed data (pickle, joblib, CSV, Parquet)

Mastering these topics will give you a strong grasp on preparing and understanding data for any ML or data science project!