**1. Dataset Exploration**

Inspect dataset shape, data types, summary statistics, and initial visualizations.

Check for correlations, missing data patterns, and class imbalance.

**2. Advanced Missing Data Handling**

Identify missing values with heatmaps, missingno library, or patterns across columns.

Impute using:

Mean/Median/Mode

KNNImputer

Regression imputation

Time-series forward/backward fill (for sequential data)

Explain why each method was chosen.

**3. Outlier Detection and Treatment**

Detect outliers using Z-score, IQR, Mahalanobis distance, or Isolation Forest.

Decide whether to remove, truncate, or transform them using log/square root/Box-Cox/Yeo-Johnson.

**4. Data Consistency Checks**

Identify inconsistent or impossible entries ( negative ages, duplicate IDs).

Correct inconsistencies and standardize categorical values.

**5. Duplicate Handling and Index Validation**

Remove duplicate rows or reset index as needed.

Ensure unique identifiers exist where required.

**6. Advanced Feature Validation**

Check cross-column logic (start ≤ end dates, total = sum of components).

Correct or flag inconsistent rows.

**7. Feature Scaling**

* Apply **Min-Max Scaling**, **Z-score Standardization**, and **Robust Scaler** to different features.
* Compare effects on distribution and variance.

**8. Handling Skewed Features**

* Identify skewed numeric features using histograms and skewness metrics.
* Apply **Log, Square Root, Box-Cox, Yeo-Johnson, or Quantile Transformation** as appropriate.

**9. Encoding Categorical Features**

* Apply **One-Hot Encoding** for nominal variables.
* Apply **Ordinal/Label Encoding** for ordinal features.
* Apply **Target Encoding or Frequency Encoding** for high-cardinality categorical features.

**10. Feature Engineering: Polynomial & Interaction Terms**

* Create **polynomial features** (degree 2 or 3) for numeric columns.
* Generate **interaction terms** to capture combined effects.

**11. Domain-Specific Feature Creation**

* Engineer **new features** based on domain knowledge.
* Example:
  + For housing: total area = bedrooms × average room size
  + For Titanic: family size = SibSp + Parch + 1

**12. Date/Time Feature Extraction**

* Extract components: year, month, day, weekday, or time differences.
* Create lag or rolling features for time-series datasets.

**13. Binning / Discretization**

* Bin continuous variables into meaningful categories (e.g., age groups, income brackets).
* Apply **KBinsDiscretizer** or custom binning logic.

**14. Feature Transformation for ML**

* Apply **Power Transformer**, **Quantile Transformer**, or **Robust Scaling** to prepare features for modeling.

**15. Handling Imbalanced Classes**

* Analyze class distribution in target variable.
* Apply **SMOTE**, **Random Oversampling**, or **undersampling** if necessary.

**16. Feature Reduction / Selection (Optional)**

* Apply **PCA** or correlation-based feature selection to reduce dimensionality.
* Justify which features are retained and why.

**17. Truncation & Winsorization**

* Limit extreme values for sensitive numeric features.
* Compare model impact before and after truncation.

**18. Automated Validation Rules**

* Implement rules to detect anomalies (e.g., negative sales, impossible temperatures).
* Flag or correct invalid rows programmatically.