1. Data Cleaning:

- **Purpose:** To handle missing values, noisy data (errors), and inconsistencies.
- **Techniques:**
  - **Missing Data:**
    - **Ignore the tuple (row):** If a row has many missing values, and the dataset is large, you might just remove that row.
      - *Example:* In a customer dataset, if a row is missing age, address, *and* purchase history, it might be discarded.
    - **Fill missing values:**
      - **Manually:** Use domain knowledge to fill in the blanks. *Example:* If you know a customer's missing "city" is likely "New York" based on their zip code, you might enter it.
      - **Using a global constant:** Use values such as "unknown", "null".
      - **Using a central tendency**: Use the attribute mean (for numerical data) or mode (for categorical data). *Example:* Fill missing "age" values with the average age of all customers.
      - **Most probable value:** Use a model (like a regression or decision tree) to predict the most likely value. *Example:* Predict a missing "income" based on the customer's "education" and "occupation."
  - **Noisy Data:**
    - **Binning:** Sort data and divide it into "bins" (groups). Smooth the values within each bin by the bin mean, median, or boundaries. *Example:* Group ages into bins like 1-10, 11-20, 21-30, etc., and replace each age with the average age of its bin.
    - **Regression:** Fit a regression model to the data to smooth out noise. *Example:* If you have "sales" vs. "advertising spend," a regression line can help smooth out daily fluctuations to show the overall trend.
    - **Clustering:** Group similar data points into clusters. Outliers (data points far from any cluster) can be considered noise. *Example:* Cluster customers based on purchasing habits; a customer with very unusual purchases might be an outlier.

2. Data Transformation:

- **Purpose:** To convert data into a suitable format for mining.
- **Techniques:**
  - **Normalization:** Scale data to a specific range (e.g., -1.0 to 1.0 or 0.0 to 1.0). *Example:* Convert all income values to a range between 0 and 1, so that income doesn't disproportionately influence a model compared to, say, age.
    Min-max normalization:

z-score normalization
Normalization by decimal scaling

- **Attribute Selection (Feature Construction):** Create new attributes from existing ones. *Example:* From "date of birth," create a new attribute "age group" (e.g., "Young Adult," "Middle-Aged," "Senior").
- **Discretization:** Convert continuous attributes into discrete (categorical) ones. *Example:* Convert "temperature" into "Cold," "Mild," "Hot."
- **Concept Hierarchy Generation:** Replace low-level attributes with higher-level concepts. *Example:* Replace "city" with "state" or "country."
- **Smoothing:** remove noise using binning, clustering and regression.
- **Aggregation:** data summarization. for example daily sales data can be aggregated to calculate monthly sales.
- **Generalization:** low-level attributes are replaced by higher level. for example, street can be generalized to city.

3. Data Reduction:

- **Purpose:** To reduce the volume of data while preserving (as much as possible) its analytical value. This is *crucial* for very large datasets.
- **Techniques:**
  - **Data Cube Aggregation:** Create a data cube by aggregating data at different levels of a hierarchy. *Example:* Aggregate daily sales data into monthly, quarterly, and yearly totals.
  - **Attribute Subset Selection:** Remove irrelevant or redundant attributes. *Example:* If "customer ID" is not useful for predicting purchase behavior, remove it. Statistical tests (p-values, significance levels) can guide this.
  - **Numerosity Reduction:**
    - **Parametric (e.g., Regression):** Assume a model (like linear regression) and store only the model parameters instead of the entire dataset. *Example:* If sales data fits a linear regression model well, store the slope and intercept instead of all individual data points.
    - **Non-parametric (e.g., Histograms, Clustering, Sampling):**
      - *Histograms:* Group data into bins and store bin counts.
      - *Clustering:* Represent data by cluster centroids.
      - *Sampling:* Select a representative subset of the data.
  - **Dimensionality Reduction:**
    - **Wavelet Transforms:** Useful for signal processing and image compression. Store a subset of wavelet coefficients.

  - **Principal Component Analysis (PCA):** Find new, uncorrelated variables (principal components) that capture most of the variance in the data. Keep only the most important components.

Key Considerations and Examples:

- **Data Quality:** The goal of preprocessing is to improve data quality (accuracy, completeness, consistency, timeliness, believability, interpretability).
- **Schema Integration:** Handling different schemas such as when merging databases, you need to resolve naming inconsistencies ("cust_id" vs. "customer_number").
- **Redundancy Detection:** Correlation analysis can identify redundant attributes (e.g., if "age" and "date of birth" are both present, "age" is redundant).
- **Tuple Duplication:** Remove duplicate rows.
- **Data Conflict Resolution:** Resolve inconsistencies in data values from different sources (e.g., different units of measurement).

By applying these preprocessing steps, you create a cleaner, more manageable, and more informative dataset that is ready for effective data mining and analysis. The specific techniques used will depend on the nature of the data and the goals of the analysis.