Data Preprocessing

Equi-Depth/Distance Binning (Bin Means)

1. Divide data into equal-sized bins.
2. Replace all values in a bin with the mean of the bin.

Example: Data: $[3.95, 4.12, 5.13, 5.84, 6.63, 7.92, 8.14, 8.78, 9.21, 10.88, 11.11, 12.12]$

- **Bins (3 bins)**:
  - Bin 1: $[3.95, 4.12, 5.13, 5.84]$, Mean = 4.76
  - Bin 2: $[6.63, 7.92, 8.14, 8.78]$, Mean = 7.87
  - Bin 3: $[9.21, 10.88, 11.11, 12.12]$, Mean = 10.83

Smoothed Data: $[4.76, 4.76, 4.76, 4.76, 7.87, 7.87, 7.87, 7.87, 10.83, 10.83, 10.83, 10.83]$

Equi-Width/Frequency Binning

1. Calculate bin width: $Width = \frac{max - min}{number\ of\ bins}$
2. divide the range of width
3. assign each data to its corresponding mean of bin

Bin Boundaries

1. Divide data into bins.
2. Replace each value with the nearest boundary of its bin.

Example: Data: $[3.95, 4.12, 5.13, 5.84]$

- **Bin Boundaries**:
  - Min = 3.95, Max = 5.84
  - Smoothed Data: $[3.95, 3.95, 5.84, 5.84]$

Normalization Techniques

Min-max normalization

Scales data to a fixed range $[0, 1]$:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Example: Original Data: $[10, 20, 30]$, Min = 10, Max = 30.
Normalized Data: $[0, 0.5, 1.0]$

Z-score Normalization

Scales data using mean (μ) and standard deviation (σ):

$$z = \frac{x - \mu}{\sigma}$$

Example: Original Data: $[10, 20, 30]$, Mean = 20, Std Dev = 10.
Normalized Data: $[-1, 0, 1]$

Handle Missing Data

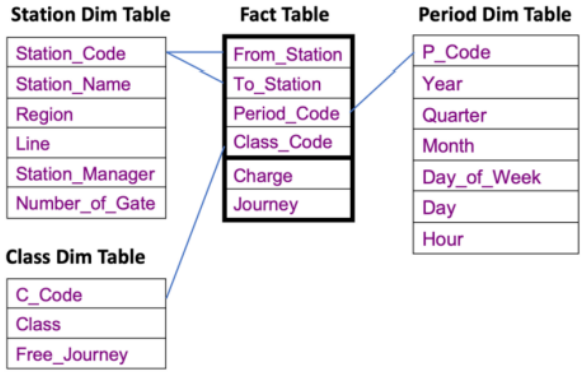
1. Fill mean/median/mode
2. Linear Interpolation: estimate values based on surrounding values [mean of surrounding numbers]
3. Association Analysis

Data Warehousing

Star Schemas

Fact Table: Contains numeric data & foreign keys to dimensions.
Dimension Tables: Contains attributes for dimensions (e.g., time, location).

Star Schema:



Scenario: A retail store records sales data.

- **Fact Table**:
  - Measures: `Sales_Amount`, `Quantity`.
  - Foreign Keys: `Product_ID`, `Time_ID`, `Store_ID`.
- **Dimension Tables**:
  1. **Product**: `Product_ID`, `Product_Name`, `Category`.
  2. **Time**: `Time_ID`, `Year`, `Month`, `Day`.
  3. **Store**: `Store_ID`, `Store_Name`, `Location`.

Data Cubes

Example:

Dimensions: `Product`, `Time`, `Store`.

Operations:

1. **Roll-Up**:
   - Aggregate data from a lower level to a higher level (e.g., daily to monthly sales).
2. **Drill-Down**:
   - Break data into finer levels (e.g., yearly sales to monthly sales).
3. **Slice**:
   - Select data for a specific dimension (e.g., sales for "Product A").
4. **Dice**:
   - Select a subset of data (e.g., sales for "Product A" in "Store X").

Roll-Up

Drill Down

Task:

- Use the **data cube** from part (c) to demonstrate **Roll-Up** and **Drill-Down**.

Explanation:

1. **Roll-Up**:
   - Aggregate data to a higher level of hierarchy.
   - Example: Aggregate daily patient data to monthly totals.
     - From `Day` → `Month` in the **Time** dimension.
2. **Drill-Down**:
   - Break data into finer levels of detail.
   - Example: Analyze regional trends by drilling down from `Region` to `City` in the **Location** dimension.

Use Case:

- Query: "Find the number of fever cases in Region A in 2023."
  - **Roll-Up**: Aggregate
  - **Drill-Down**: Focus on Region A → Cities within Region A.

Data Preprocessing

Clean and transform raw data for better analysis.
Cleaning: Handle missing values, outliers, noise.
Integration: Combine data from multiple sources.
Reduction: Dimensionality reduction, sampling.
Discretization: Transform continuous data into intervals.
Importance: Quality data is essential for meaningful results

Feature Engineering

- Enhance input data to improve model performance.

Techniques:

- Encoding: Convert categorical data (e.g., one-hot encoding).
- Imputation: Handle missing values (e.g., mean, mode).
- Transformation: Scale/normalize numerical data.
- New Features: Create derived variables, extract features from text/images.

Poor feature engineering limits even the best algorithms

Web Mining

Types:

1. Web Content Mining: Extract useful information from web pages.
2. Web Structure Mining: Analyze hyperlink structures.
3. Web Usage Mining: Understand user behavior.

Key Techniques:

1. Cosine similarity for document matching.
2. Keyword-based retrieval (e.g., stop words, stemming).
3. Clustering for text categorization

Data Warehousing Features

- **Subject-Oriented**: Organized by subject (e.g., customer, sales).
- **Integrated**: Data consistency across sources.
- **Time-Variant**: Stores historical data.
- **Non-Volatile**: Data updates do not overwrite existing data.