Data Warehousing and Data Mining (DWDM) Concepts

# 1. Different Types of Attributes with Examples

\*\*Attributes\*\* are the characteristics or properties of data in a dataset. Here are the different types:

1. \*\*Nominal Attributes:\*\* Categories with no specific order.   
 - Example: Colors (Red, Blue, Green), Gender (Male, Female)

2. \*\*Ordinal Attributes:\*\* Categories with a clear order.   
 - Example: Movie ratings (Poor, Fair, Good, Excellent), Education level (High School, Bachelor's, Master's, Ph.D.)

3. \*\*Interval Attributes:\*\* Numeric values with meaningful intervals, but no true zero point.   
 - Example: Temperature (20°C, 30°C), Dates (Years like 1990, 2000, 2010)

4. \*\*Ratio Attributes:\*\* Numeric values with a true zero, allowing for comparison of magnitudes.   
 - Example: Height (160 cm, 180 cm), Salary ($30,000, $60,000)

5. \*\*Discrete Attributes:\*\* Countable values.   
 - Example: Number of students in a class (20, 25, 30), Number of cars in a parking lot (10, 15, 20)

6. \*\*Continuous Attributes:\*\* Values that can take any number within a range.   
 - Example: Time (10.5 seconds, 20.7 seconds), Temperature (25.3°C, 30.6°C)

7. \*\*Cyclic Attributes:\*\* Values that repeat in a cycle.   
 - Example: Days of the week (Monday, Tuesday), Seasons (Spring, Summer, Fall, Winter)

# 2. Different Data Sets of Attributes with Examples

\*\*Data sets\*\* refer to the collection of data points, and each data point is characterized by attributes. Here are some types:

1. \*\*Univariate Data Set:\*\* Data set with only one attribute.   
 - Example: List of students' ages [12, 13, 14, 15]

2. \*\*Bivariate Data Set:\*\* Data set with two attributes.   
 - Example: Height vs. Weight [(160 cm, 55 kg), (170 cm, 65 kg)]

3. \*\*Multivariate Data Set:\*\* Data set with more than two attributes.   
 - Example: Sales data [(Product A, $30, 20 units), (Product B, $50, 15 units)]

4. \*\*Time Series Data Set:\*\* Data set where each data point is collected over time.   
 - Example: Monthly sales figures [January: $1000, February: $1200, March: $1500]

5. \*\*Spatial Data Set:\*\* Data set with spatial attributes.   
 - Example: Geographic coordinates [(Latitude: 34.05, Longitude: -118.25), (Latitude: 40.71, Longitude: -74.01)]

# 3. Preprocessing in Detail

\*\*Data preprocessing\*\* is the process of preparing raw data for analysis. It involves several steps:

1. \*\*Data Cleaning:\*\* Removing noise and correcting inconsistencies.  
 - Techniques: Handling missing values, removing duplicates, correcting errors.

2. \*\*Data Integration:\*\* Combining data from multiple sources into a coherent dataset.  
 - Techniques: Data merging, data matching, data consolidation.

3. \*\*Data Transformation:\*\* Converting data into a suitable format or structure for analysis.  
 - Techniques: Normalization, aggregation, encoding categorical attributes.

4. \*\*Data Reduction:\*\* Reducing the volume of data while maintaining its integrity.  
 - Techniques: Dimensionality reduction (e.g., PCA), data compression, sampling.

5. \*\*Data Discretization:\*\* Converting continuous attributes into discrete ones.  
 - Techniques: Binning, histogram analysis.

# 4. Sales Price Records Binning

Given the sales price records: \*\*5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215\*\*

## a) Equal-Frequency (Equal-Depth) Partitioning

Divide the data into bins so that each bin has an equal number of records.  
- \*\*Bin 1:\*\* 5, 10, 11, 13  
- \*\*Bin 2:\*\* 15, 35, 50, 55  
- \*\*Bin 3:\*\* 72, 92, 204, 215

## b) Equal-Width Partitioning

Divide the range of data into intervals of equal width.  
- \*\*Bin 1 (5-75):\*\* 5, 10, 11, 13, 15, 35, 50, 55, 72  
- \*\*Bin 2 (76-145):\*\* 92  
- \*\*Bin 3 (146-215):\*\* 204, 215

# 5. Data Transformation & Data Discretization

\*\*Data Transformation\*\*: Converting data into a suitable format for analysis.   
- Techniques: Normalization, Standardization, Aggregation, Encoding.

\*\*Data Discretization\*\*: Converting continuous attributes into discrete intervals.   
- Techniques: Binning, Histogram Analysis, Clustering, Decision Trees.