Data Preprocessing

1. Data Preprocessing Overview

Data Quality : Why we preprocess the data

- Definition: Data preprocessing involves preparing raw data for analysis by improving its quality, consistency, and usability.
- Importance :
  - Real-world data is often noisy, incomplete and inconsistent.
  - Effective preprocessing ensures reliable and accurate analysis.

Key Concepts in Data Quality

1. Dimensions of data quality

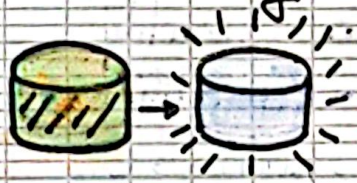
Dimension	Description	Example
Accuracy	Correctness of data	Age '20' instead of '200'
Completeness	Availability of all values	Missing marital status in records
Consistency	Uniformity in representation	Dates: "2023/11/23" vs. "23-11-2023"
Timeliness	Data is up to date	Old addresses not updated
Believability	Trustworthiness of data	Survey data manipulated
Interpretability	Easy comprehension	Using Clear variable names

2. Examples of Data Issues

- **Noise and Outliers:**
  Error or extreme values (e.g. a salary of "-10" or "10M")
- **Missing Values:**
  Gaps in attributes information
- **Duplicate Data:**
  Repeated entries with minor variations (e.g. same customer entered twice with slightly different names)

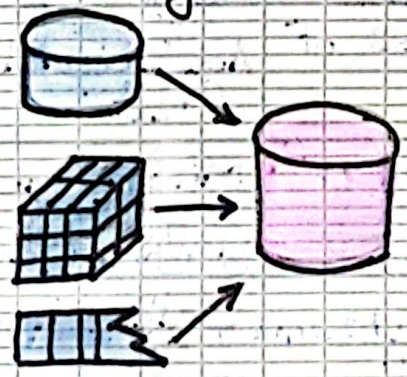
↳ Major Tasks in Data Preprocessing

Data Cleaning



- fill in missing values
- Smooth noisy data
- Identify or remove outliers, and resolve inconsistencies

Data Integration



- Integration of multiple databases, data cubes, or files.

Data Transformation

- -2, 32, 100, 59 → -0.02, 0.32, 1.00, 0.59 e.g normalizatio

Data Reduction

	A1	A2	A3 ...	A126
T1				
T2				
T3				
T4				
...				
T2000				

	A1	A3 ...	A115
T1			
T4			
...			
T1456			

- Dimensionality Reduction.

2. Data Cleaning

What is Data Cleaning?

Data Cleaning is the process of detecting and correcting (or removing) errors and inconsistencies in data to improve its quality.

Common Data Issues and Solutions

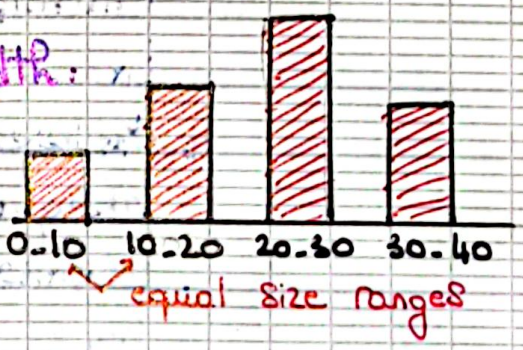
1. Incomplete Data:
   - Difinition: Missing attribute values or incomplete records.
   - Examples:
     - Missing "Occupation" field in a survey
     - Missing rows in financial transactions.
   - How to handle:
     - Ignore the Tuple:
       - Discard (delete) incomplete records
       - Cons: Inefficient if many records are incomplete
     - Manual Filling:
       - Requires domain knowledge
       - Cons: Time-Consuming and error-prone.
     - Automatic Filling:
       - Global Constant: Assign a placeholder (e.g. "unknown")
       - Central Tendency Measures: fill with mean, median, or mode.
       - Inference: Predict missing values using ML (e.g. Bayesian methods, decision Trees.)

2. Noisy Data

- Definition: Data containing random errors or outliers
- Examples:
  - Age recorded as "-5" or "150"
  - Price with extreme values (e.g. $1M for an ordinary item)
- How to handle:
  - Binning:
    - Definition: data smoothing technique, used to handle noisy data by grouping (or partitioning) data into intervals (bins) and replacing individual values with representative values.
    - Types of Binning Methods:
      1) Equal-Width Binning: (= distance)
         - Def: Divides the range of data into bins of equal size
         - Steps:
           - Calculate the bin width:

$$\text{Bin Width} = \frac{\text{Max Val.} - \text{Min Val.}}{\text{Nb. of bins}}$$

           - Assign values to bins based on their range



         - Example:
           - Dataset: [4, 8, 9, 15, 21, 24] (Sorted)
           - Nb. of bins: 3
           - Min = 4, Max = 24 → Bin Width = (24-4)/3 = 6.67.
           - Bins:
             Bin 1: [4 - 10.67] → contains 4, 8, 9
             Bin 2: [10.67 - 17.33] → contains 15
             Bin 3: [17.33 - 24] → contains 21, 24

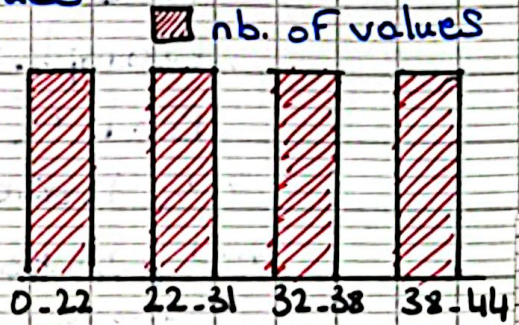
2) Equal-Depth Binning (= Frequency)

. Def: Divides data into bins containing approx. the same nb. of values

. Steps:

→ Sort the data in asc. order

→ Divide into N bins, each containing roughly Total Values/N values



. Example:

. Dataset = [4, 8, 9, 15, 21, 24] (Sorted)

. Nb. of bins : 3

. Bins :

Bin 1 : [4, 8]

Bin 2 : [9, 15]

Bin 3 : [21, 24]

. Smoothing Techniques for Bins

1) Smoothing by Bin Means

. Replace all values in a bin with the mean value of that bin

. Example:

Bin 1 : [4, 8, 9]

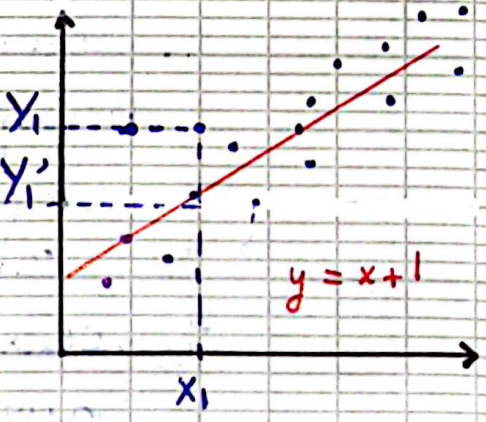
↳ Mean = (4+8+9)/3 = 7

↳ Smoothed Bin : [7, 7, 7]

2) Smoothing By Bin Median
- Replace all values in a bin with the median value of that bin
- Example:
  Bin 1 : [4, 8, 9]
  ↳ Median: 8
  ↳ Smoothed Bin : [8, 8, 8]

3) Smoothing By Bin Boundaries
- Replace values near a boundary with the nearest boundary value
- Example
  Bin 1 : [4, 8, 9, 15]
  ↳ lower boundary = 4, upper boundary = 15
  ↳ Smoothed Bin : [4, 4, 4, 15]

↳ Regression
- Fit the data to a regression model
- Replace noisy or missing values by predicted values
- Requires model of attributes dependencies
- Can be used for both data smoothing and handling missing values.



↳ Clustering
. Group similar values together and treat outliers as noise.

3. Inconsistent Data
. Def: Discrepancies in coding, naming, or formatting.
. Examples:
↳ Different formats: "2023-11-23" vs. "23/11/2023"
↳ Duplicate records with minor variations.
. How to handle:
↳ Identify Discrepancies
. Use domain knowledge or metadata to define acceptable formats
. Apply statistical methods to detect outliers or dependencies.
↳ Correct Errors:
. Manually correct if feasible
. Apply transformations to standardize formats

↳ Data Cleaning as a Process.
1. Identify data discrepancies
. Sources:
Poorly designed data entry forms.
Human error in data entry.

. How to identify data discrepancies?
→ Use domain knowledge, metadata
→ what are the data type and domain name of each attribute?
→ what are the acceptable values for each attribute?
→ Use basic statistical data descriptions (mean, mode, median; symmetric vs. skewed, find outliers ...)
→ Find inconsistent use of code and any inconsistent data representations.

2. Correct data inconsistencies
. Apply transformations or manual corrections

3. Iterate
Some corrections may introduce new inconsistencies requiring iteration.

3. Data Integration
→ What is Data Integration?
. Data integration involves combining data from multiple sources into a unified dataset or data warehouse, ensuring coherence, consistency and usability.
. Purpose:
Create a holistic view of data.
Facilitate analysis by resolving discrepancies between datasets.

Challenges in Data Integration

1. Entity Identification Problem
   - Identifying the same real-world entity across different datasets
   - Example: "Bill Clinton" in one dataset may be "William Clinton" in another.
   - Solution: Use metadata or domain-specific rules to resolve inconsistencies.

2. Schema Integration
   - Aligning attribute names and data formats across datasets
   - Example:
     - Dataset A: Customer_ID, Dataset B: Cust_ID
     - Dataset A: Dates in DD/MM/YYYY
       Dataset B: Dates in MM/DD/YYYY
   - Solution: Standardize schema attributes.

3. Redundancy
   - Definition: Repeated or derived info across datasets
   - Example: Annual Income derived from Monthly Income x12
   - Solution: Detect redundancies using correlation analysis or covariance.

4. Correlation and Dependency
   - Some attributes may have strong correlations causing redundancy
   - Techniques: Statistical methods (correlation coefficient, χ^2 Test)

Techniques for handling Redundancy

1. Correlation Analysis (Nominal Data)

- Uses χ^2 (Chi-Square) Test to evaluate the dependency between categorical attributes
- Steps:
  - Create a contingency table summarizing frequencies
  - Calculate expected frequencies
  - Compute χ^2 statistic:

$$\chi^2 = \sum \frac{(\text{observed} - \text{Expected})^2}{\text{Expected}}$$

where:
observed values: the actual count
Expected values: the count obtained from contingency table. joint events
  - Compare χ^2 to a critical value from the Chi-Square table
- Example:

	dog	Cat	bird	total
men	207	282	241	730
women	234	242	232	708
total	441	524	473	1438

Contingency table Summarizing the relationships between gender and buying different type of pets

The aim of the test is to conclude whether the 2 variables (gender and choice of pet) are related to each other.

Null hypothesis Ho: there is no relation between the variables

↳ IF x^2 <= critical value, then Ho holds true.

. Table of calculated (expected) values
(row total x column total) / grand total

	dog	cat	bird	total
men	223.873	266.008	240.118	730
women	217.126	257.991	232.881	708
total	441	524	473	1438

. Chi Square Table
(observed value - calculate value)^2/calculated value

expected values

Contingency Table values ↳

observed (o)	calculated (c)	$(o-c)^2/c$
207	223.873	1.271757
282	266.008	0.96137
⋮	⋮	⋮
Total	$x^2 = 4.5422282698$	

← Chi Square formula

. Find Critical Value of Chi-Square
- degree of Freedom for the dataset:
(nb. of rows - 1) x (nb. of columns - 1)
= (2 - 1) x (3 - 1) = 2

. The tabular or critical value of chi-Square for 2 degrees of freedom and 0.05 significance factor is 5.991

. 5.991 > 4.54
↳ critical value of x^2 > = calculated value of x^2

. So H_0 is accepted, which mean the variables do not have a significant relationship.

2. Correlation Coefficient (Numerical Data)

. Measures the linear relationship between 2 numerical attributes

. Formula:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1) \cdot \sigma_x \cdot \sigma_y}$$

. $r \in [-1, 1]$, where:
↳ $r = 1$: perfect positive correlation
↳ $r = 0$: No correlation
↳ $r = -1$: perfect negative correlation

4. Data Transformation

→ What is Data Transformation?

. A function that maps the entire set of attribute values to a new set, ensuring that each old value corresponds to a new one

→ Methods in Data Transformation

1) Smoothing

. Removes noise from data to enhance quality

2) Aggregation
- Combines two or more attributes/objects into one
- Purpose
  - ↳ Reduce the nb. of attributes/objects
  - ↳ Alters scales (e.g. aggregating cities into regions)

3) Normalization
- Scales data into a specified range
- Techniques:
  - ↳ Min-Max Normalization: Scales values to a range like [0,1]
  - Formula: $v' = \frac{(v - min_A)}{(max_A - min_A)} \cdot (new\_max_A - new\_min_A) + new\_min_A$
  - Example: Income range $12,000 - $98,000 normalized to [0,1].
    $73,600 is mapped as:
    $$v' = \frac{(73,600 - 12,000)}{(98,000 - 12,000)} \cdot (1.0 - 0.0) + 0.0$$
    $$= 0.716$$
  - ↳ Z-Score Normalization: Centers data using the mean (μ) and standard deviation
  - Formula: $v' = \frac{(v - \mu)}{\sigma}$
  - Example: If $\mu = 54,000$, $\sigma = 16,000$ and $v = 73,600$
    $$v' = \frac{(73,600 - 54,000)}{16,000} = 1.225$$

↳ Normalization by Decimal Scaling: Scales values by powers of 10(j) until the largest absolute value becomes < 1

Formula: $v' = \frac{v}{10^j}$, where j: smallest in such that Max(|v'|) < 1

5. Data Reduction

→ What is Data Reduction?

The process of obtaining a smaller representation of the data that preserves its integrity for analysis.

→ Strategies for Data Reduction

1) Dimensionality Reduction

Eliminates irrelevant or redundant attributes

Techniques:

↳ Principal Component Analysis (PCA)
Identifies patterns and compress data into principal components while retaining variance

↳ Singular Value Decomposition (SVD)
Decomposes data into matrices for simplification

↳ Wavelet Transform:
Used for image compression and data summarization

2) Numerosity Reduction

Represent data with fewer rows or summarized models

- Techniques:
  - Regression models: Fit data to functions
  - Log-linear models: Summarize large, cat. data
  - Histograms: divide data into intervals and summarize with counts
  - Clustering: Group similar data points and represent each group with a centroid
  - Sampling: Select a subset of data points for analysis
  - Data Cube Aggregation: Summarize data across multiple dimensions (e.g. from daily to monthly sales)

3) Data Compression
- Encodes data more efficiently to save space
- Examples:
  - Huffman Encoding
  - Lossy compression techniques for image and audio.