Data Processing and Visualization

Redwan Ahmed Rizvee
https://rizveeredwan.github.io

Contents

- Data object and attribute types: nominal, binary, ordinal, numeric
- Basic statistical descriptions of data
- Measuring data similarity and dissimilarity
- Data preprocessing: cleaning, integration, transformation, and discretization
- Data visualization techniques:
  - Pixel-oriented
  - Geometric projection
  - Icon-based
  - Hierarchical visualization

Data object and attribute types

A **data object** represents an entity (e.g., a person, a product, a transaction) in the dataset. Each object is described using a set of **attributes** (also known as features, variables, or fields).

Example:

For a student object:

{Name: "Alice", Gender: "Female", Grade: "A", Age: 21}

Here, each piece of information is an attribute.

Attribute Types

a. Nominal

- **Definition**: **Categorical data with no inherent order**.
- **Examples**:
  - Gender: Male, Female
  - Color: Red, Blue, Green
  - Department: HR, Sales, Engineering

b. Binary

- **Definition**: Special case of nominal attribute with only **two categories**.
- **Types**:
  - **Symmetric**: Both values are equally important.
    Example: isMarried = {Yes, No}
  - **Asymmetric**: One value is more important.
    Example: hasDisease = {Yes, No} (Yes is more significant)
- **Values**: Often encoded as 0 and 1

Attribute Types

c. Ordinal

- **Definition**: Categories that have a **meaningful order**, but no fixed distance between them.
- **Examples**:
  - Education level: High School < Bachelor < Master < PhD
  - Customer satisfaction: Low, Medium, High
- **Operations**: Can rank but not do arithmetic on theme

d. **Numeric (Quantitative)**

Divided into two subtypes:

- **Interval**: Ordered, meaningful differences, but no true zero
  → Example: Temperature in Celsius (20°C - 10°C = 10°C, but 0°C ≠ no temperature)
- **Ratio**: Ordered, meaningful differences, and **true zero** exists
  → Example: Age, Height, Weight, Salary (0 = absence)

Basic statistical descriptions of data

1. Measures of Central Tendency

These describe the **center or average** of the data.

- **Mean (Average)**:
  Mean= $\sum x_i$ / N; Sum of all values divided by the number of values.
- **Median**:
  The middle value when data is sorted. For even-sized data, it's the average of the two middle values.
- **Mode**:
  The most frequently occurring value in the dataset.

Basic statistical descriptions of data

2. Measures of Dispersion (Spread)

These describe how much the data varies.

- **Range**: Range=Max−Min
- **Variance**:
  Average of the squared differences from the mean.
  Variance=$\sum(x_i-\bar{x})^2/n$ (for population) and Variance = $\sum(x_i-\bar{x})^2/(n-1)$ (for sample)

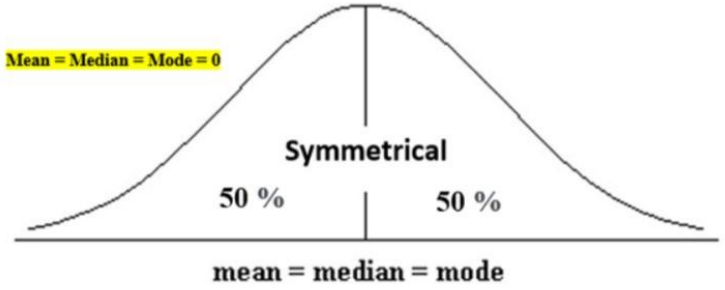
  This denominator correction is done to correct the biasness, known as Bessel's correction

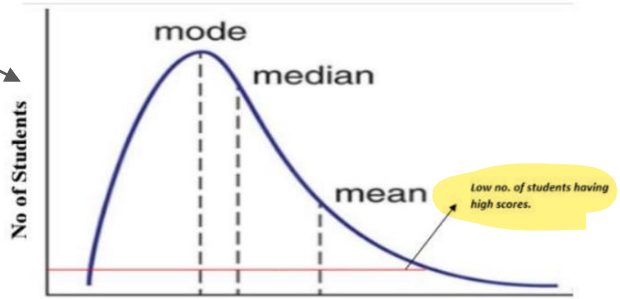
- **Standard Deviation (SD)**:
  Square root of the variance; gives dispersion in the same unit as the data.
- **Interquartile Range (IQR)**:
  IQR=Q3−Q1, Q1 (First Quartile) = Median of the data between (0-50%) and Q3 (Third Quartile) = Median of the data between (50-100)%
  Measures the spread of the middle 50% of the data.

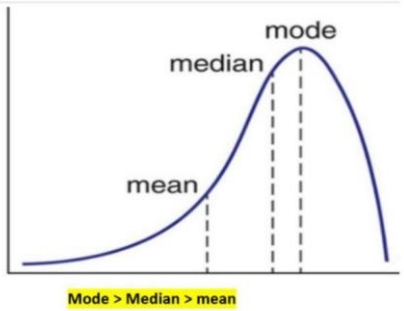
Basic statistical descriptions of data

3. Shape of the Distribution

- **Skewness**:
  Indicates asymmetry of the data distribution.
  - Positive skew: tail on the right
  - Negative skew: tail on the left
- **Kurtosis**:
  Measures "tailedness" or the presence of outliers.
  - High kurtosis: heavy tails (outliers)
  - Low kurtosis: light tails



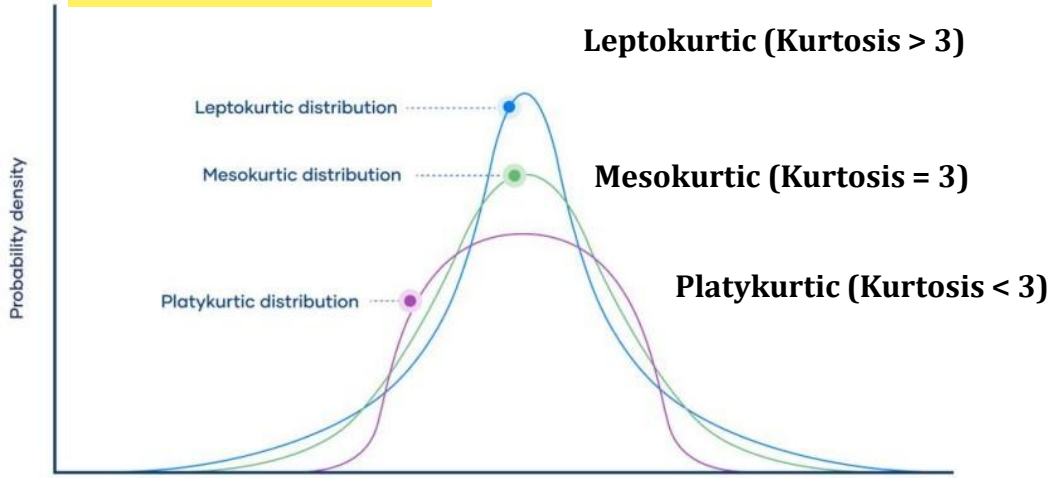




Basic statistical descriptions of data

- **Kurtosis**:

Kurtosis is a statistical measure that **quantifies the shape of a probability distribution**. It provides information about the **tails** and **peakedness of the distribution compared to a normal distribution**. **Positive kurtosis** indicates **heavier tails** and a **more peaked distribution**, while **negative kurtosis** suggests **lighter tails** and **a flatter distribution**. Kurtosis helps in **analyzing the characteristics** and **outliers** of a dataset.



Source

Measuring data similarity and dissimilarity

Similarity: A numerical measure of how alike two data objects are. Higher values mean more similarity.

Dissimilarity: A numerical measure of how different two data objects are. Higher values mean more dissimilarity.

Numerical Data

- **Euclidean Distance** (most common):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- **Manhattan Distance** (L1 norm):

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

- **Minkowski Distance** (general form):

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$

- **Cosine Similarity**:

$$\text{cosine}(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

Used especially in high-dimensional data (e.g., text vectors)

Distance Measures

Euclidian: continuous numerical data, same scale or have been normalized

Manhattan: numerical data with outliers, **avoid squaring differences** due to creating more gaps

Minkowski: generalized distance function that includes both Euclidean (p=2) and Manhattan (p=1), **tune the sensitivity** to large differences by adjusting ppp

Data is **high-dimensional and sparse** (many zeros).

Distance Measures

Categorical Data

- **Simple Matching Coefficient (SMC):**

$$\text{SMC} = \frac{\text{Number of matching attributes}}{\text{Total attributes}}$$

- **Jaccard Coefficient** (ignores mutual 0s):

$$\text{Jaccard}(x, y) = \frac{|x \cap y|}{|x \cup y|}$$

SMC: categorical or binary features, **Presence and absence** are equally important, **symmetric** comparison

Jaccard: categorical or binary features, consider **presence more than absence, Asymmetric features**

Distance Measures

Ordinal Data

- Convert to ranks, then use numeric distance measures (e.g., Euclidean on ranks).

Binary Data

- Symmetric Binary (e.g., gender):
  - Use Hamming Distance or Jaccard.

- Asymmetric Binary (e.g., disease presence):
  - Use Jaccard Coefficient.

Example: Jaccard Similar

Example (Binary Case)

Two binary vectors:

- A = [1, 0, 1, 1, 0]
- B = [1, 1, 0, 1, 0]

We interpret them as sets of positions where value is 1:

- A = {1, 3, 4}
- B = {1, 2, 4}

Then:

- $A \cap B = \{1,4\} \rightarrow$ size = 2
- $A \cup B = \{1,2,3,4\} \rightarrow$ size = 4

Jaccard Similarity=2/4=0.5, Jaccard Distance=1−0.5=0.5

Example (Set-based Example)

- A = {"apple", "banana", "cherry"}
- B = {"banana", "cherry", "date", "fig"}

Then:

- $A \cap B = \{\text{"banana","cherry"}\} \rightarrow$ size = 2
- $A \cup B = \{\text{"apple","banana","cherry","date","fig"}\} \rightarrow$ size = 5

Jaccard Similarity=2/5=0.4
Jaccard Distance=1−0.4=0.6

* Jaccard Only Focuses only on **positive** matches

Data Preprocessing: Cleaning

1. Data Cleaning

 Purpose:

To remove **inaccuracies**, **inconsistencies**, **missing values**, and **noisy data**.

 Common Problems:

- **Missing values** (e.g., blank cells in a CSV)
- **Noisy data** (e.g., typos, outliers)
- **Inconsistent formatting** (e.g., "BD", "Bangladesh", "bangla desh")
- **Duplicate entries**

 Techniques:

- **Handling missing values**:
  - Ignore the record (if rare)
  - Fill with mean/median/mode
  - Use interpolation or ML models
- **Noise reduction**:
  - Smoothing (e.g., binning, regression)
  - Outlier detection
- **Standardization**:
  - Ensure uniform formats (dates, case, units)
- **Deduplication**:
  - Identify and remove duplicate rows

Data Integration

2. Data Integration

 Purpose: To combine data from **multiple sources** into a **coherent dataset**.

 Scenarios:

- Merging data from: Multiple databases, Internal and external systems, IoT devices + business logs

 Challenges:

- **Schema integration**: Same concept may use different names (e.g., "Cust_ID" vs "CustomerID")
- **Entity resolution**: Matching same real-world entity (e.g., "John Smith" vs "J. Smith")
- **Redundancy**: Repeated data from different sources
- **Data conflicts**: Conflicting values for the same attribute

 Techniques:

- Use **metadata** and **domain knowledge** for schema alignment
- Apply **record linkage / entity matching algorithms**
- Use **ETL tools** for scalable integration

Data Transformation

Purpose:

To convert data into a **suitable format** for analysis or mining.

Common Methods:

- **Normalization/Standardization**: Scale numerical values to a common range (e.g., [0,1] or z-scores)
- **Attribute construction**: Derive new meaningful features (e.g., Age = CurrentYear - BirthYear)
- **Encoding categorical variables**: One-hot encoding, label encoding
- **Aggregation**: Summarizing data (e.g., total monthly sales from daily data)
- **Smoothing**: Reduce noise via averaging or regression
- **Generalization**: Replace low-level data with high-level concepts (e.g., “25” → “Young Adult”)

Data Discretization

Purpose:

To convert **continuous** attributes into **categorical** ones.

Why Needed:

- Some models (e.g., decision trees) perform better with categorical input.
- To simplify patterns or enhance interpretability.

Methods:

- **Equal-width binning**: Divide range into intervals of equal size
- **Equal-frequency binning**: Each bin has the same number of records
- **Cluster-based discretization**: Use clustering (e.g., k-means) to form bins
- **Entropy-based** (e.g., in decision trees): Discretization based on class information gain

Data visualization techniques: Pixel-oriented

Every **data value** is mapped to a **colored pixel** on the screen. Very useful for **large-scale** multidimensional data. Uses color intensity or hue to encode data values. Efficient for datasets with millions of data points.

Techniques: **Recursive Pattern, Circle Segments, Spiral Pixel Displays, etc.**

Example:

Heatmap visualization using seaborn

Data Visualization Technique: Geometric Projection

Projects **high-dimensional data** onto a **2D or 3D geometric space** using mathematical transformations. Preserves structure such as **clusters**, **distances**, or **relationships**.

Techniques: Scatter Plots (for 2D/3D), **Parallel Coordinates, Principal Component Analysis (PCA), t-SNE**, **UMAP** (nonlinear dimensionality reduction)

Scatterplot visualization using seaborn

Data Visualization Technique: Icon-based

Each **data item** is represented by a **small icon**, where icon features (shape, size, color, angle, etc.) encode attribute values.

Techniques: **Chernoff Faces** (facial features represent data dimensions), **Star Glyphs** or **Radar Charts** (spokes or arms encode attributes), **Stick Figures**

Data Visualization Technique: Icon-based

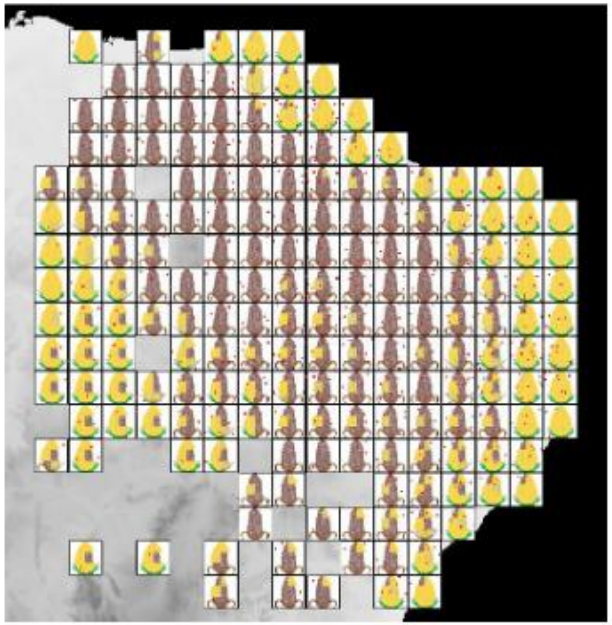


Figure 6. Mosaic image in regular layout with stations faded in (red dots)

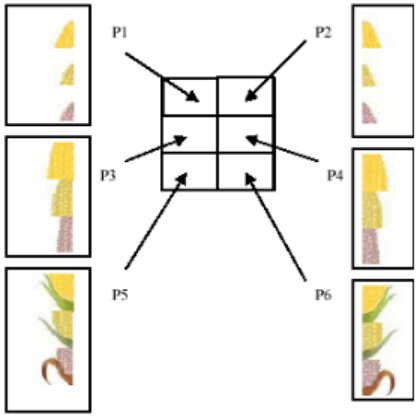


Figure 3. Construction of a metaphor-based icon, representing six parameters

Source

Figure 2. The three base icons displaying maize conditions: good (left), middle (center) and bad (right) conditions

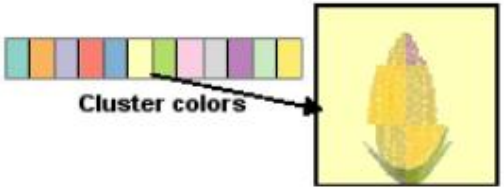


Figure 4. Metaphor-based icon representing six parameters with the background color identifying the cluster

Data Visualization Technique: Hierarchical Visualization

Designed to represent **hierarchical (tree-like or nested) data structures**.

- **Techniques:** Tree Maps, Dendrograms, Radial Trees, Sunburst Charts

Tree map visualization using plotly

Dendogram visualization using plotly

Questions?