Course: Data Mining
Lecturer: Hend Muslim Jasim **Date: November ,2025**
Reference: " Data Mining Concepts and Techniques" Third Edition, Jiawei Han, et al.

Lecture 2: Knowing Data and Its Types

1. Introduction to Data Objects and Attributes

In data mining, we analyze **data sets** composed of **data objects**. A data object represents an entity such as a customer, product, or patient and is described by **attributes**.

- **Data Object**: Also known as a *sample*, *instance*, *record*, or *tuple*.
- **Attribute**: A data field representing a characteristic of a data object (e.g., *age*, *height*, *price*). Also called *variable*, *feature*, or *dimension*.

2- Types of Attributes:

Attributes can be classified into the following types:

1. **Nominal**
2. **Binary**
3. **Ordinal**
4. **Numeric**

1- Nominal Attributes

- **Definition**: Attributes whose values are symbols or names representing categories. The values have no meaningful order.
- **Examples**:
  - hair_color: black, brown, blond, red
  - marital_status: single, married, divorced
  - occupation: teacher, engineer, doctor
- **Properties**:
  - Values are categories or states.
  - Mathematical operations are not meaningful.
  - Mode (most frequent value) is a useful measure.

2- Binary Attributes

- **Definition**: Nominal attributes with only two possible states: 0 or 1.
- **Examples**:
  - smoker: yes (1) or no (0)
  - medical_test: positive (1) or negative (0)
- **Types**:
  - **Symmetric**: Both states are equally important (e.g., *gender*).
  - **Asymmetric**: One state is more important than the other (e.g., *HIV test result*).

3- Ordinal Attributes

- **Definition**: Attributes with values that have a meaningful order or ranking, but the differences between values are not known.
- **Examples**:
  - drink_size: small, medium, large
  - education_level: high school, bachelor, master, PhD
  - customer_rating: poor, fair, good, excellent
- **Properties**:
  - Order is meaningful.
  - Differences between values are not quantifiable.
  - Median and mode are appropriate measures of central tendency.

4- Numeric Attributes

- **Definition**: Quantitative attributes represented by integer or real values.

- **Types**:
  - **Interval-Scaled**:
    - Measured on a scale with equal units.
    - Differences are meaningful, but ratios are not.
    - **Example**: Temperature in Celsius or Fahrenheit.
  - **Ratio-Scaled**:
    - Has a true zero point.
    - Ratios between values are meaningful.
    - **Examples**: Height, weight, age, income.

5- Discrete vs. Continuous Attributes

- **Discrete Attribute**:
  - Has a finite or countably infinite set of values.
  - **Examples**: customer_ID, zip_code, number_of_children.
- **Continuous Attribute**:
  - Has real numbers as values.
  - **Examples**: height, weight, temperature.

6- Basic Statistical Descriptions of Data

To understand data, we use **statistical measures**:

a. Measures of Central Tendency

- **Mean**: The average value.
- **Median**: The middle value in a sorted list.
- **Mode**: The most frequently occurring value.
- **Midrange**: The average of the largest and smallest values.

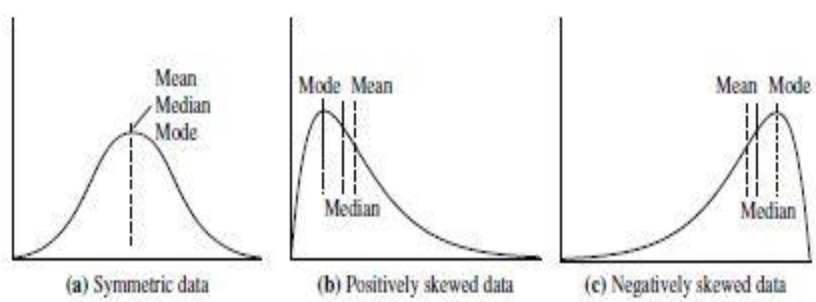

Figure 2.1 Mean, median, and mode of symmetric versus positively and negatively skewed data.

b. Measures of Data Dispersion

- **Range**: Difference between max and min values.
- **Quartiles**: Q1 (25th percentile), Q2 (median), Q3 (75th percentile).
- **Interquartile Range (IQR)**: IQR = Q3 – Q1.

Course: Data Mining
Lecturer: Hend Muslim Jasim **Date: November ,2025**
Reference: " Data Mining Concepts and Techniques" Third Edition, Jiawei Han, et al.

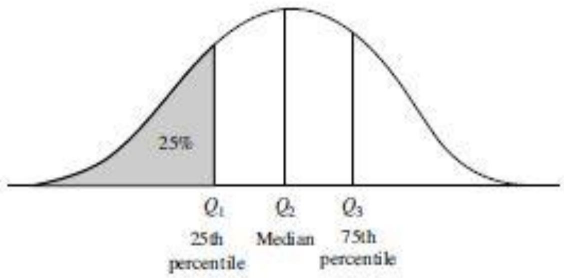

Figure 2.2 A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

- **Variance & Standard Deviation**: Measures of data spread.
- **Five-Number Summary:** Min, Q1, Median, Q3, Max.

c. Graphic Displays

- **Boxplots**: Visualize five-number summary and outliers.

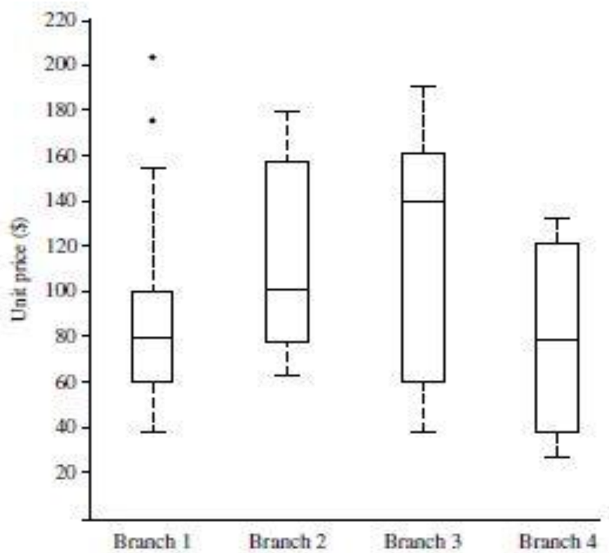

Figure 2.3 Boxplot for the unit price data for items sold at four branches of *AllElectronics* during a given time period.

- **Histograms**: Show frequency distributions.

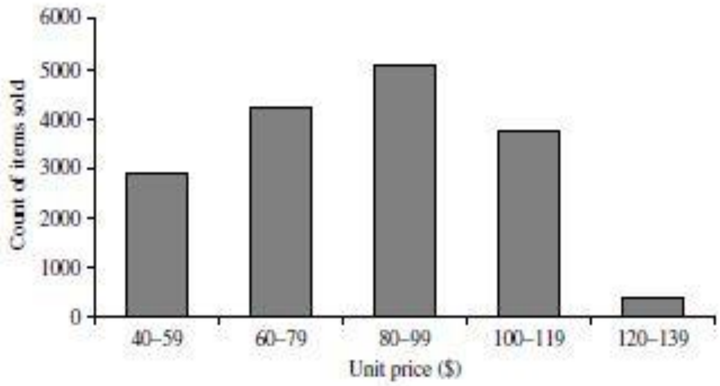

Figure 2.6 A histogram for the Table 2.1 data set.

- **Scatter Plots**: Display relationships between two numeric attributes.

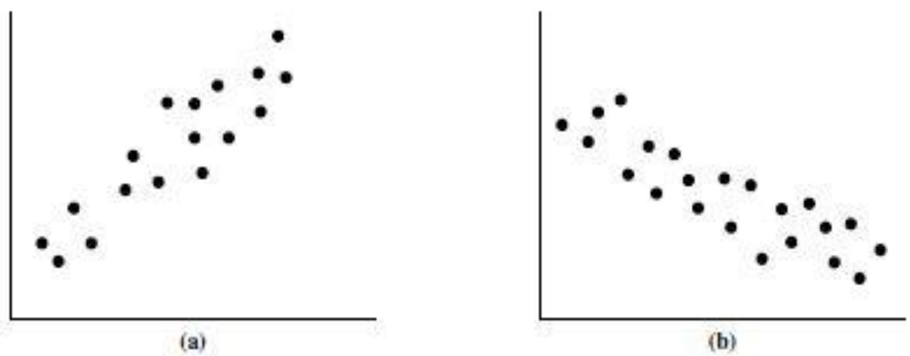

Figure 2.8 Scatter plots can be used to find (a) positive or (b) negative correlations between attributes.

- **Quantile Plots**: Compare distributions.

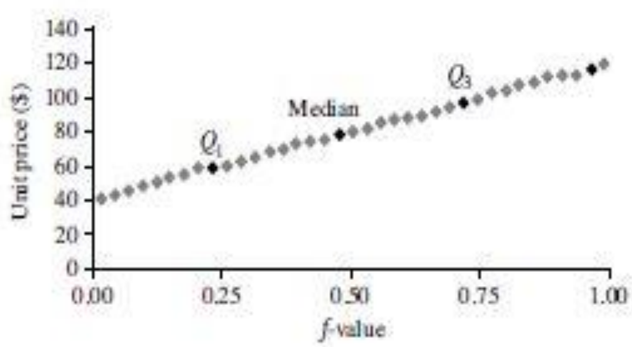

Figure 2.4 A quantile plot for the unit price data of Table 2.1.

Course: Data Mining
Lecturer: Hend Muslim Jasim **Date: November ,2025**
Reference: " Data Mining Concepts and Techniques" Third Edition, Jiawei Han, et al.

8. Data Visualization Techniques

Visualization helps in understanding data patterns and relationships.

a. Pixel-Oriented Techniques

- Each attribute value is represented by a colored pixel.

Figure 2.10 Pixel-oriented visualization of four attributes by sorting all customers in *income* ascending order.

b. Geometric Projection Techniques

- **Scatter Plots**: 2-D or 3-D plots of data points.

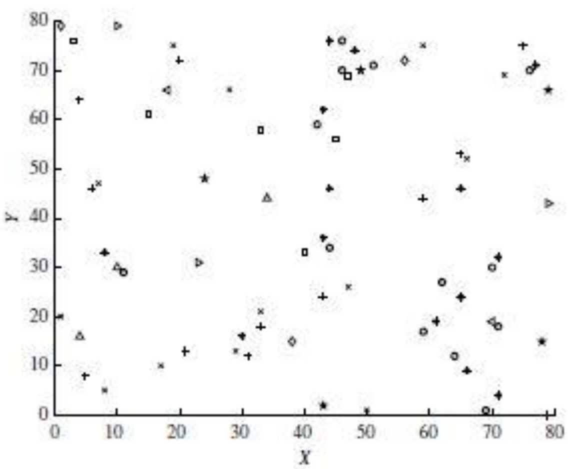

Figure 2.13 Visualization of a 2-D data set using a scatter plot. *Source: www.cs.sfu.ca/jpei/publications/rareevent-geoinformatica06.pdf*.

- **Parallel Coordinates**: Each attribute is represented by a parallel axis.

c. Icon-Based Techniques

- **Chernoff Faces**: Represent multi-dimensional data using facial features.

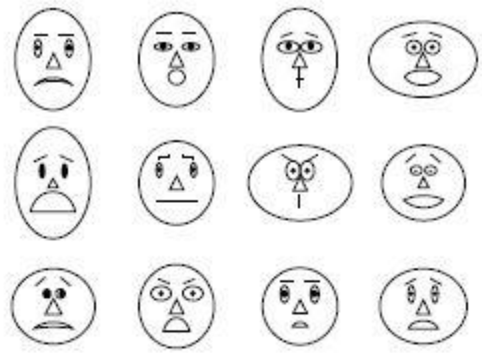

Figure 2.17 Chernoff faces. Each face represents an *n*-dimensional data point ($n \leq 18$).

- **Stick Figures**: Use limb angles and lengths to represent attributes.

Figure 2.18 Census data represented using stick figures. *Source:* Professor G. Grinstein, Department of Computer Science, University of Massachusetts at Lowell.

d. Hierarchical Techniques

- **Tree-maps**: Display hierarchical data as nested rectangles.

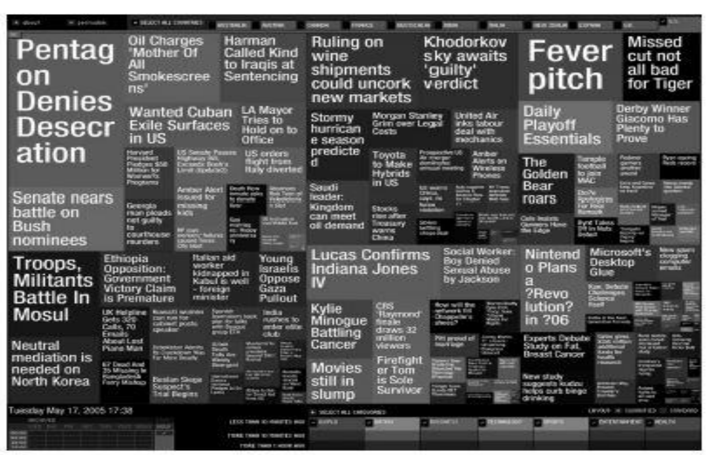

Figure 2.20 Newsmap: Use of tree-maps to visualize Google news headline stories. *Source: www.cs.umd.edu/class/spring2005/cmsc838s/viz4all/ss/newsmap.png.*

- **Worlds-within-Worlds (n-Vision)**: Explore high-dimensional data in multiple levels.

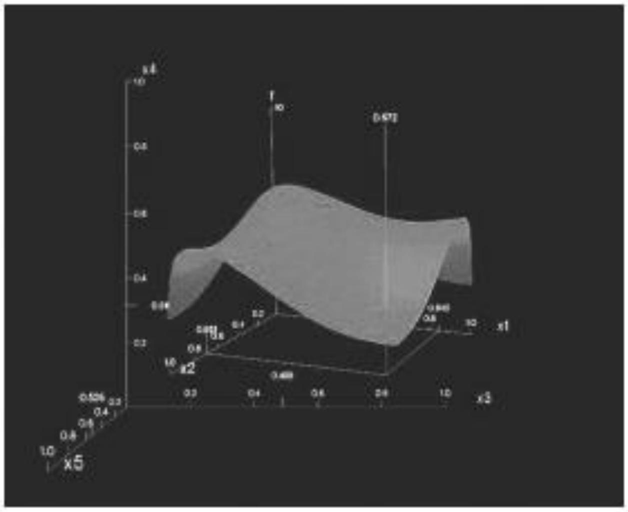

Figure 2.19 "Worlds-within-Worlds" (also known as *n*-Vision). *Source: http://graphics.cs.columbia.edu/projects/AutoVisual/images/1.dipstick.5.gif.*

9. Measuring Data Similarity and Dissimilarity

In many data mining tasks (e.g., clustering, classification), we need to measure how similar or dissimilar objects are.

a. Data Structures

- **Data Matrix**: n×p*n*×*p* matrix storing n*n* objects with p*p* attributes.

Data matrix (or *object-by-attribute structure*): This structure stores the *n* data objects in the form of a relational table, or *n*-by-*p* matrix (*n* objects × *p* attributes):

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}. \qquad (2.8)$$

- **Dissimilarity Matrix**: n×n*n*×*n* matrix storing pairwise dissimilarities.

  - **Dissimilarity matrix** (or *object-by-object structure*): This structure stores a collection of proximities that are available for all pairs of *n* objects. It is often represented by an *n*-by-*n* table:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ \vdots & \vdots & \vdots & & \\ d(n,1) & d(n,2) & \cdots & \cdots & 0 \end{bmatrix}, \tag{2.9}$$

  where $d(i, j)$ is the measured **dissimilarity** or "difference" between objects i and j. In general, $d(i, j)$ is a non-negative number that is close to 0 when objects i and j are highly similar or "near" each other, and becomes larger the more they differ. Note that $d(i, i) = 0$; that is, the difference between an object and itself is 0. Furthermore, $d(i, j) = d(j, i)$. (For readability, we do not show the $d(j, i)$ entries; the matrix is symmetric.) Measures of dissimilarity are discussed throughout the remainder of this chapter.

b. Proximity Measures by Attribute Type

- **Nominal Attributes**:
  - d(i,j)=p−mp*d*(*i*,*j*)=*pp*−*m*, where m*m* is the number of matches.
- **Binary Attributes**:
  - **Symmetric**: d(i,j)=r+sq+r+s+t*d*(*i*,*j*)=*q*+*r*+*s*+*tr*+*s*
  - **Asymmetric (Jaccard Coefficient)**: d(i,j)=r+sq+r+s*d*(*i*,*j*)=*q*+*r*+*sr*+*s*
- **Numeric Attributes**:
  - **Euclidean Distance**:
    d(i,j)=∑k=1p(xik−xjk)2*d*(*i*,*j*)=*k*=1∑*p*(*xik*−*xjk*)2
  - **Manhattan Distance**:
    d(i,j)=∑k=1p|xik−xjk|*d*(*i*,*j*)=*k*=1∑*p*|*xik*−*xjk*|
  - **Minkowski Distance**:
    d(i,j)=(∑k=1p|xik−xjk|h)1/h*d*(*i*,*j*)=(*k*=1∑*p*|*xik*−*xjk*|*h*)1/*h*
- **Ordinal Attributes**:
  - Replace values with ranks, normalize, then use numeric distance measures.
- **Mixed Types**:
  - Combine dissimilarities from different attribute types into a single measure.
- **Cosine Similarity** (for text or sparse data):
  - sim(x,y)=x·y‖x‖‖y‖*sim*(*x*,*y*)=‖*x*‖‖*y*‖*x*·*y*

Summary

- Understanding **data types** is essential for choosing the right data mining techniques.
- **Statistical descriptions** and **visualization** help explore data characteristics.
- **Similarity and dissimilarity measures** are crucial for clustering, classification, and outlier detection.

By knowing your data and its types, you can preprocess it effectively and apply appropriate mining methods to extract meaningful patterns.