

Chapter I: Introduction & Foundations

▶ 1.1 Introduction

- 1.1.1 Definitions & Motivations
- 1.1.2 Data to be Mined
- 1.1.3 Knowledge to be discovered
- 1.1.4 Techniques Utilized
- 1.1.5 Applications Adapted
- 1.1.6 Major Issues in Data Mining

▶ 1.2 Getting to Know Your Data

- 1.2.1 Data Objects and Attribute Types
- 1.2.2 Measuring Data Similarity and Dissimilarity
- 1.2.3 Descriptive Data Summarization

▶ 1.3 Basics from Probability Theory and Statistics

- 1.3.1 Probability Theory
- 1.3.2 Statistical Inference: Sampling and Estimation
- 1.3.3 Statistical Inference: Hypothesis Testing and Regression

1.2.1 Data Objects and Attribute Types

- ▶ Types of data sets
- ▶ Data objects
- ▶ Attributes and their types

Types of Data Sets

▶ Record

- Relational records
- Data matrix, e.g., numerical matrix, cross tabulations.
- Document data: text documents: term-frequency vector
- Transaction data

Relational records

Login	First name	Last name
koala	John	Clemens
lion	Mary	Stevens

} record

Login	phone
koala	039689852639

Transactional data

TID	Items
	Books
1	Bred, Cake, Milk
2	Beer, Bred

} record

Document data

	team	ball	lost	timeout
Document1	3	5	2	2
Document2	0	0	3	0
Dccument3	0	1	0	0

} record

Cross tabulation

	Books	Multimedia devices
Big spenders	30%	70%
Budget spenders	60%	25%
Very Tight spenders	10%	5%

} record

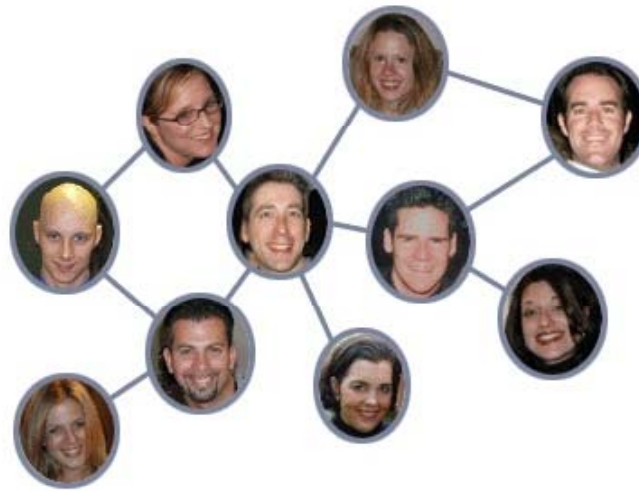
Types of Data Sets

▶ Graph and Network

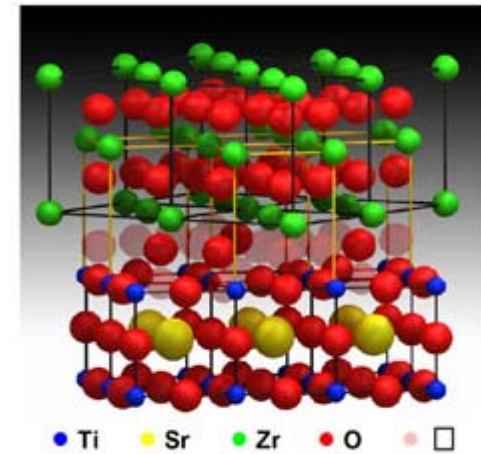
- World Wide Web
- Social or information networks
- Molecular structures networks



World Wide Web



Social Networks



Molecular Structures Network

Types of Data Sets

▶ Ordered

- Videos
- Temporal data
- Sequential data
- Genetic sequence data

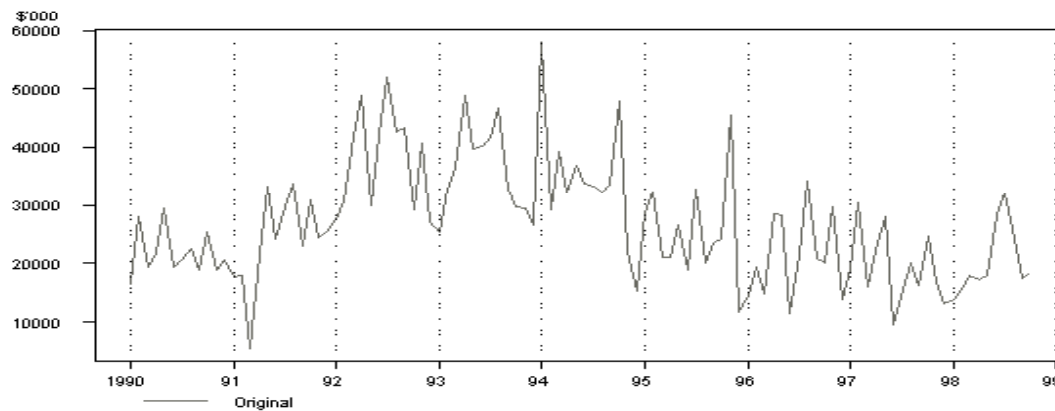


Video: sequence of mages

Transactional sequence

Computer-> Web cam ->USB key

Generic Sequence: DNA-code



Temporal data: Time-series
monthly Value of Building Approvals

Types of Data Sets

▶ Spatial, image and multimedia

- Spatial data
- Image data
- Video data
- Audio Data



Spatial data: maps



Images



Videos



Audios

Data Objects

- ▶ Data sets are made up of data objects.
- ▶ A **data object** represents an entity.
- ▶ **Examples**
 - **Sales database**: customers, store items, sales
 - **Medical database**: patients, treatments
 - **University database**: students, professors, courses
- ▶ Also called *samples*, *examples*, *instances*, *data points*, *objects*, *tuples*.
- ▶ Data objects are described by **attributes**.
- ▶ Database rows -> data objects; columns -> attributes.

Patient_ID	Age	Height	Weight	Gender
1569	30	1,76m	70 kg	male
2596	26	1,65m	58kg	female

Data Object

Attributes

Attributes

- ▶ **Attribute** (or **dimension**, **feature**, **variable**) is a data field, representing a characteristic or a feature of a data object.
→ E.g., *Patient_ID*, *name*, *address*

Patient_ID	Name	Age	Height	Weight	Gender
1569	John	30	1,76m	70 kg	male
2596	Mary	26	1,65m	58kg	female

Attributes

- ▶ Types
 - Nominal
 - Binary
 - Numeric: quantitative
 - Interval-scaled
 - Ratio-scaled

Attribute Types

- ▶ **Nominal** categories, states, or “**names of things**”
 - $Hair_color = \{black, brown, blond, red, grey, white\}$
 - marital status, occupation, ID numbers, zip codes
- ▶ **Binary**
 - Nominal attribute with only 2 states (**0 and 1**)
 - Symmetric binary: both outcomes equally important
 - e.g., gender
 - Asymmetric binary: outcomes not equally important.
 - e.g., medical test (**positive vs. negative**)
 - Convention: assign 1 to most important outcome (e.g., HIV positive)
- ▶ **Ordinal**
 - Values have a meaningful order (**ranking**) but magnitude between successive values is not known.
 - $Size = \{small, medium, large\}$, grades, army rankings

Numeric Attributes Types

- ▶ Quantity (integer or real-valued)
- ▶ **Interval-Scaled**
 - Measured on a scale of equal-sized units
 - Values have order
 - No true zero-point
 - E.g., temperature in C° or F°, calendar dates
 - we can add and subtract degrees **-100° is 10° warmer than 90°**, we cannot multiply values or create ratios **-100° is not twice as warm as 50°**.
- ▶ **Ratio-Scaled**
 - Inherent zero-point
 - We can speak of values as being an order of magnitude larger than the unit of measurement
 - E.g., temperature in Kelvin, length, counts, monetary quantities
 - A **6-foot** person is **20% taller** than a **5-foot** person.
 - A baseball game lasting **3 hours** is **50%** longer than a game lasting **2 hours**.

Discrete vs. Continuous Attributes

▶ Discrete Attribute

- Has only a finite or countably infinite set of values
 - E.g., zip codes, profession, or the set of words in a collection of documents
- Sometimes, represented as integer variables
- Note: Binary attributes are a special case of discrete attributes

▶ Continuous Attribute

- Has real numbers as attribute values
 - E.g., temperature, height, or weight
- Practically, real values can only be measured and represented using a finite number of digits
- Continuous attributes are typically represented as floating-point variables(float, double , long double)

Chapter I: Introduction & Foundations

▶ 1.1 Introduction

- 1.1.1 Definitions & Motivations
- 1.1.2 Data to be Mined
- 1.1.3 Knowledge to be discovered
- 1.1.4 Techniques Utilized
- 1.1.5 Applications Adapted
- 1.1.6 Major Issues in Data Mining

▶ 1.2 Getting to Know Your Data

- 1.2.1 Data Objects and Attribute Types
- 1.2.2 Measuring Data Similarity and Dissimilarity
- 1.2.3 Descriptive Data Summarization

▶ 1.3 Basics from Probability Theory and Statistics

- 1.3.1 Probability Theory
- 1.3.2 Statistical Inference: Sampling and Estimation
- 1.3.3 Statistical Inference: Hypothesis Testing and Regression

1.2.2 Data Similarity and Dissimilarity

▶ **Similarity**

- Numerical measure of how alike two data objects are
- Value is higher when objects are more alike
- Often falls in the range $[0,1]$

▶ **Dissimilarity** (e.g., distance)

- Numerical measure of how different two data objects are
- Lower when objects are more alike
- Minimum dissimilarity is often 0
- Upper limit varies

▶ **Proximity** refers to a similarity or dissimilarity

Data Matrix and Dissimilarity Matrix

▶ Data matrix

- n data points with p dimensions
- Two modes: rows and columns represent different entities

$$\begin{array}{c} \text{Data} \\ \text{objects} \end{array} \begin{array}{c} \text{Variables} \\ \left[\begin{array}{ccccc} x_{11} & \dots & x_{1f} & \dots & x_{1p} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{if} & \dots & x_{ip} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nf} & \dots & x_{np} \end{array} \right] \end{array}$$

▶ Dissimilarity matrix

- n data points, but registers only the distance

$$\begin{array}{c} \text{Data} \\ \text{objects} \end{array} \begin{array}{c} \text{Data} \\ \text{objects} \\ \left[\begin{array}{cccc} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ \vdots & \vdots & \vdots & \\ d(n,1) & d(n,2) & \dots & \dots & 0 \end{array} \right] \end{array}$$

Nominal Attributes

- ▶ Can take 2 or more states, e.g., red, yellow, blue, green (generalization of a binary attribute)
- ▶ **Method 1:** Simple matching
 - m : # of matches, p : total # of variables

$$d(i, j) = \frac{p - m}{p}$$

- ▶ **Method 2:** Use a large number of binary attributes
 - creating a new binary attribute for each of the M nominal states

Binary Attributes

- ▶ A contingency table for binary data

		Object j		
		1	0	sum
Object i	1	q	r	$q+r$
	0	s	t	$s+t$
sum		$q+s$	$r+t$	p

- ▶ Distance measure for symmetric binary variables

$$d(i, j) = \frac{r + s}{q + r + s + t}$$

- ▶ Distance measure for asymmetric binary variables

$$d(i, j) = \frac{r + s}{q + r + s}$$

- ▶ Jaccard coefficient (*similarity* measure for *asymmetric* binary variables)

$$\text{sim}(i, j) = \frac{q}{q + r + s} = 1 - d(i, j)$$

Binary Attributes

▶ Example

Name	Gender	Fever	Cough	Test-1	Test-2	Test-3	Test-4
Jack	M	Y	N	P	N	N	N
Mary	F	Y	N	P	N	P	N
Jim	M	Y	P	N	N	N	N

- gender is a symmetric attribute
- the remaining attributes are asymmetric binary
- let the values Y and P be set to 1, and the value N be set to 0

$$d(\text{jack}, \text{mary}) = \frac{0 + 1}{2 + 0 + 1} = 0.33$$

$$d(\text{jack}, \text{jim}) = \frac{1 + 1}{1 + 1 + 1} = 0.67$$

$$d(\text{jim}, \text{mary}) = \frac{1 + 2}{1 + 1 + 2} = 0.75$$

Numeric Attributes

- ▶ The measurement unit used for interval-scale attributes can have an effect on the similarity
 - E.g., kilograms vs. pounds for weight

- ▶ **Need of standardizing the data**

- Convert the original measurements to unit-less variables
- For measurements of each variable f :
 - Calculate the **mean absolute deviation**, s_f

$$s_f = \frac{1}{n} (|x_{1f} - m_f| + |x_{2f} - m_f| + \dots + |x_{nf} - m_f|)$$

m_f the mean of f
 x_{1f}, \dots, x_{nf} : measurements of f

- Calculate the standardized measurement, or z-score

$$z_{if} = \frac{x_{if} - m_f}{s_f}$$

- Using the mean absolute deviation reduces the effect of outliers
- Outliers remain detectable (non squared deviation)

Distance on Numeric Data

- ▶ **Minkowski distance**: A popular distance measure

$$d(i, j) = \sqrt[h]{(|x_{i_1} - x_{j_1}|^h + |x_{i_2} - x_{j_2}|^h + \dots + |x_{i_p} - x_{j_p}|^h)}$$

where $i = (x_{i_1}, x_{i_2}, \dots, x_{i_p})$ and $j = (x_{j_1}, x_{j_2}, \dots, x_{j_p})$ are two p -dimensional data objects, and h is the order

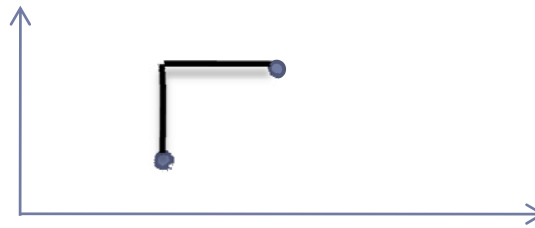
- ▶ **Properties**

- $d(i, j) > 0$ if $i \neq j$, and $d(i, i) = 0$ (Positive definiteness)
 - $d(i, j) = d(j, i)$ (Symmetry)
 - $d(i, j) \leq d(i, k) + d(k, j)$ (Triangle Inequality)
- ▶ A distance that satisfies these properties is a **metric**

Special Cases of Minkowski Distance

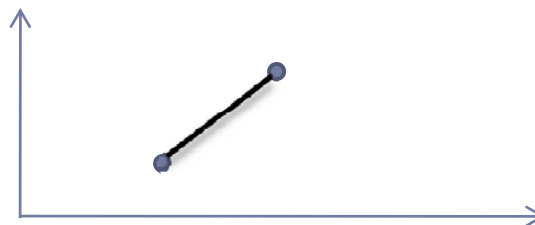
- ▶ $h = 1$: **Manhattan** (city block, L_1 norm) distance
 - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i, j) = |x_{i_1} - x_{j_1}| + |x_{i_2} - x_{j_2}| + \dots + |x_{i_p} - x_{j_p}|$$



- ▶ $h = 2$: (L_2 norm) **Euclidean** distance

$$d(i, j) = \sqrt{(|x_{i_1} - x_{j_1}|^2 + |x_{i_2} - x_{j_2}|^2 + \dots + |x_{i_p} - x_{j_p}|^2)}$$



Example: Minkowski Distance

Data Objects

point	attribute 1	attribute 2
x1	1	2
x2	3	5
x3	2	0
x4	4	5

Dissimilarity Matrices

Manhattan (L_1)

L	x1	x2	x3	x4
x1	0			
x2	5	0		
x3	3	6	0	
x4	6	1	7	0

Euclidean (L_2)

L2	x1	x2	x3	x4
x1	0			
x2	3.61	0		
x3	2.24	5.1	0	
x4	4.24	1	5.39	0

Vector Objects

- ▶ A **document** can be represented by thousands of attributes, each recording the *frequency* of a particular word (such as keywords) or phrase in the document.

<i>Document</i>	<i>teamcoach</i>	<i>hockey</i>	<i>baseball</i>	<i>soccer</i>	<i>penalty</i>	<i>score</i>	<i>win</i>	<i>loss</i>	<i>season</i>
Document1	5	0	3	0	2	0	2	0	0
Document2	3	0	2	0	1	1	1	0	1
Document3	0	7	0	2	1	0	3	0	0
Document4	0	1	0	0	1	2	0	3	0

- ▶ **Other vector objects**: gene features in micro-arrays, ...
- ▶ **Applications**: information retrieval, biologic taxonomy, gene feature mapping, ...
- ▶ **Cosine measure**: If d_1 and d_2 are two vectors (e.g., term-frequency vectors), then

$$\cos(d_1, d_2) = (d_1 \bullet d_2) / (||d_1|| \cdot ||d_2||)$$

where \bullet indicates **vector dot product**, $||\mathbf{d}||$: **length of vector d**

Cosine Similarity

- ▶ **Example:** Find the **similarity** between documents 1 and 2

$$d_1 = (5, 0, 3, 0, 2, 0, 0, 2, 0, 0)$$

$$d_2 = (3, 0, 2, 0, 1, 1, 0, 1, 0, 1)$$

- ▶ **Compute $d_1 \cdot d_2$**

$$d_1 \cdot d_2 = 5 \cdot 3 + 0 \cdot 0 + 3 \cdot 2 + 0 \cdot 0 + 2 \cdot 1 + 0 \cdot 1 + 0 \cdot 1 + 2 \cdot 1 + 0 \cdot 0 + 0 \cdot 1 = 25$$

- ▶ **Compute $\|d_1\|$**

$$\|d_1\| = (5^2 + 0^2 + 3^2 + 0^2 + 2^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2)^{0.5} = (42)^{0.5} = 6.481$$

- ▶ **Compute $\|d_2\|$**

$$\|d_2\| = (3^2 + 0^2 + 2^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2)^{0.5} = (17)^{0.5} = 4.12$$

$$\cos(d_1, d_2) = (d_1 \cdot d_2) / (\|d_1\| \|d_2\|) = 25 / (6.481 \times 4.12) = 0.94$$

Ordinal Variables

- ▶ An ordinal variable can be discrete or continuous
- ▶ Order is important, e.g., rank
- ▶ Can be treated like interval-scaled
 - replace x_{if} by their rank $r_{if} \in \{1, \dots, M_f\}$
 - map the range of each variable onto $[0, 1]$ by replacing i -th object in the f -th variable by

$$z_{if} = \frac{r_{if} - 1}{M_f - 1}$$

- compute the dissimilarity using methods for interval-scaled variables

Chapter I: Introduction & Foundations

▶ 1.1 Introduction

- 1.1.1 Definitions & Motivations
- 1.1.2 Data to be Mined
- 1.1.3 Knowledge to be discovered
- 1.1.4 Techniques Utilized
- 1.1.5 Applications Adapted
- 1.1.6 Major Issues in Data Mining

▶ 1.2 Getting to Know Your Data

- 1.2.1 Data Objects and Attribute Types
- 1.2.2 Measuring Data Similarity and Dissimilarity
- 1.2.3 Descriptive Data Summarization

▶ 1.3 Basics from Probability Theory and Statistics

- 1.3.1 Probability Theory
- 1.3.2 Statistical Inference: Sampling and Estimation
- 1.3.3 Statistical Inference: Hypothesis Testing and Regression

1.2.3 Descriptive Data Summarization

▶ Motivation

- For data preprocessing, it is essential to have an overall picture of your data
- Data summarization techniques can be used to
 - Define the typical properties of the data
 - Highlight which data should be treated as noise or outliers

▶ Data characteristics

- Central Tendency
- Data Dispersion

▶ From the data mining point of view it is important to

- Examine how these measures are computed efficiently
- Introduce the notions of **distributive** measure, **algebraic** measure and **holistic** measure

Measuring the Central Tendency

▶ Mean (algebraic measure)

Note: n is sample size

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- A **distributive** measure can be computed by partitioning the data into smaller subsets
- An **algebraic** measure can be computed by applying an algebraic function to one or more distributive measures (e.g., **mean=sum/count**)

▶ Sometimes each value x_i is weighted

- Weighted arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}$$

▶ Problem

- The mean measure is sensitive to extreme (e.g., outlier) values
- What to do?
- Trimmed mean: chopping extreme values

Measuring the Central Tendency

▶ Median (holistic measure)

- Middle value if odd number of values, or average of the middle two values otherwise
- A **holistic** measure must be computed on the entire data set
- Holistic **measures** are much more expensive to compute than **distributive** measures
- Can be estimated by interpolation (for grouped data):

$$median = L_1 + \left(\frac{n/2 - (\sum freq)l}{freq_{median}} \right) width$$

Example

<i>age</i>	<i>frequency</i>
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700

- Median interval contains the median frequency
 - **L₁**: the lower boundary of the median interval
 - **N**: the number of values in the entire dataset
 - **(Σ freq)_l**: sum of all freq of intervals below the median interval
 - **Freq_{median}** and **width** : frequency & width of the median interval

Measuring the Central Tendency

▶ Mode

- Value that occurs most frequently in the data
- It is possible that several different values have the greatest frequency: Unimodal, bimodal, trimodal, multimodal
- If each data value occurs only once then there is no mode
- Empirical formula:

$$mean - mode = 3 \times (mean - median)$$

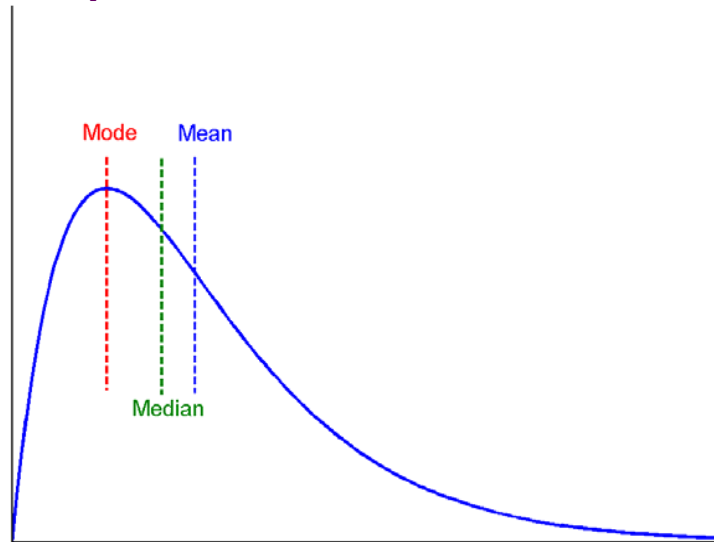
▶ Midrange

- Can also be used to assess the central tendency
- It is the average of the **smallest** and the **largest** value of the set
- It is an algebraic measure that is easy to compute

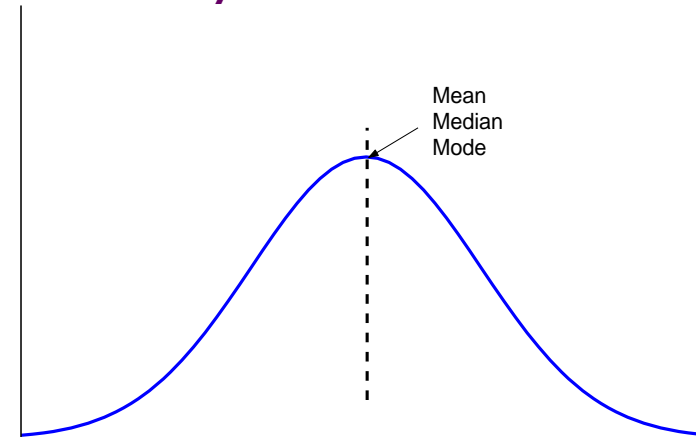
Symmetric vs. Skewed Data

- ▶ Median, mean and mode of symmetric, positively and negatively skewed data

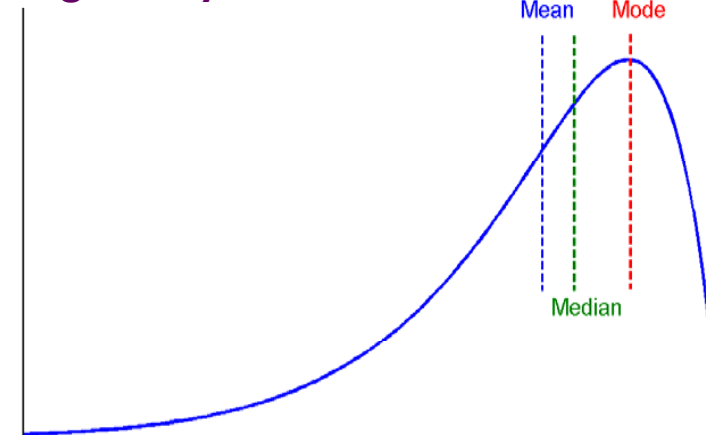
Positively skewed data



Symmetric data



Negatively skewed data



Measuring the Dispersion of Data

- ▶ The degree in which data tend to spread is called the **dispersion**, or **variance** of the data
- ▶ The most common measures for data dispersion are **range**, the **five-number summary** (based on quartiles), **the inter-quartile range**, and **standard deviation**.
- ▶ **Range**
 - The distance between the largest and the smallest values
- ▶ **Kth percentile**
 - Value x_i having the property that **k%** of the data lies at or below x_i
 - The median is **50th** percentile
 - The most popular percentiles other than the median are **Quartiles** **Q₁** (25th percentile), **Q₃** (75th percentile)
 - Quartiles + median give some indication of the center, spread, and the shape of a distribution

Measuring the Dispersion of Data

▶ Inter-quartile range

- Distance between the first and the third quartiles **$IQR=Q3-Q1$**
- A simple measure of spread that gives the range covered by the middle half of the data
- **Outlier**: usually, a value falling at least **$1.5 \times IQR$** above the third quartile or below the first quartile

▶ Five number summary

- Provide in addition information about the endpoints (e.g., tails)
- **min, Q_1 , median, Q_3 , max**
 - E.g., $\text{min} = Q1 - 1.5 \times IQR$, $\text{max} = Q3 + 1.5 \times IQR$
- Represented by a Boxplot

Measuring the Dispersion of Data

- ▶ Variance and standard deviation

- **Variance**: (algebraic, scalable computation)

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2$$

- **Standard deviation** σ is the square root of variance σ^2

- ▶ **Basic properties of the standard deviation**

- σ measures spread about the mean and should be used only when the mean is chosen as the measure of the center

- $\sigma=0$ only when there is no spread, that is, when all observations have the same value. Otherwise $\sigma>0$.

- Variance and standard deviation are **algebraic** measures. Thus, their computation is scalable in large databases.

Graphic Displays

- ▶ **Boxplot:** graphic display of five-number summary
- ▶ **Histogram:** x-axis are values, y-axis repres. frequencies
- ▶ **Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane
- ▶ **Quantile plot:** each value x_i is paired with f_i indicating that approximately $100 f_i \%$ of data are $\leq x_i$
- ▶ **Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another

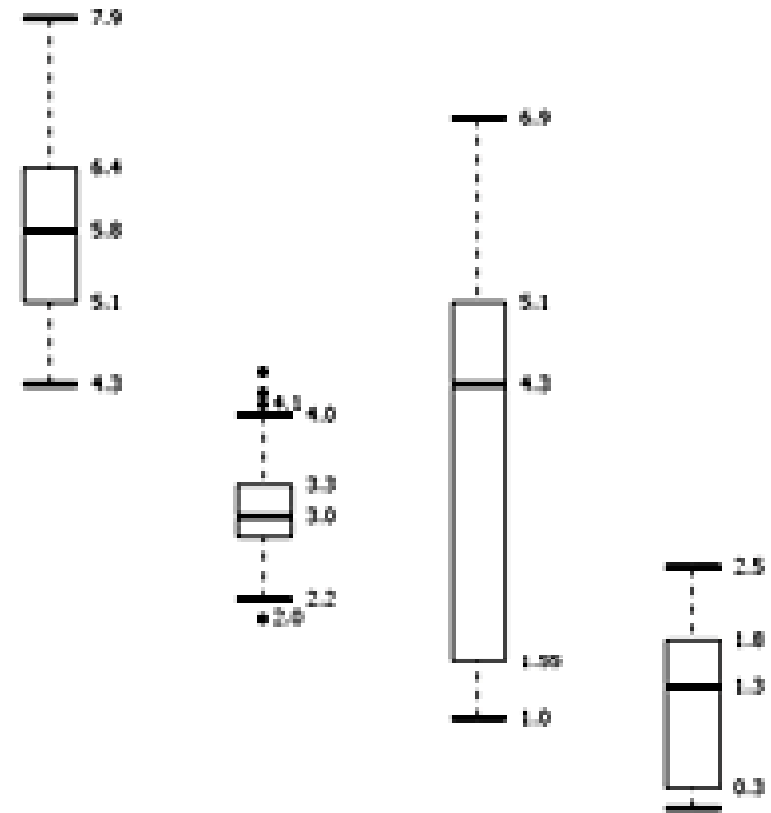
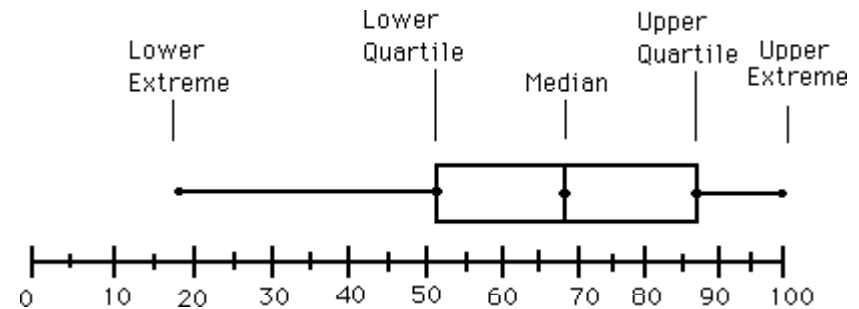
Boxplot

▶ Five-number summary of a distribution

→ Minimum, Q1, Median, Q3, Maximum

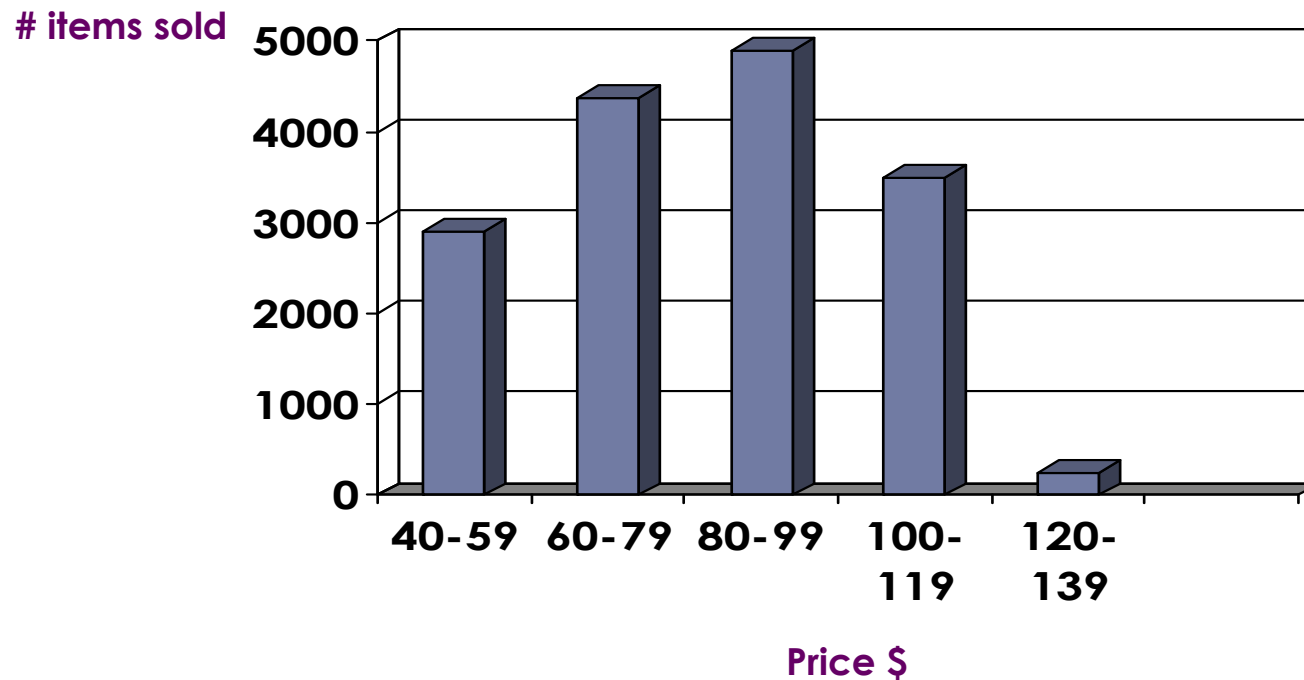
▶ Boxplot

- Data is represented with a box
- The ends of the box are at the first and third quartiles, i.e., the height of the box is IQR
- The median is marked by a line within the box
- **Whiskers**: two lines outside the box extended to Minimum and Maximum
- **Outliers**: points beyond a specified outlier threshold, plotted individually



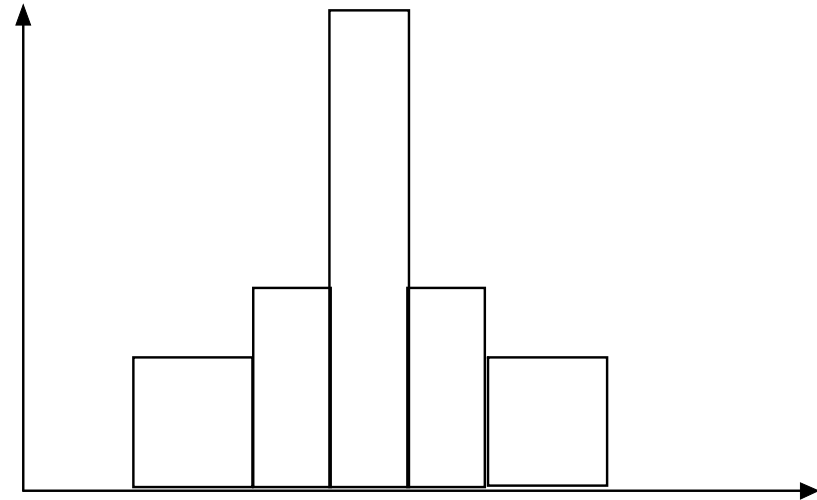
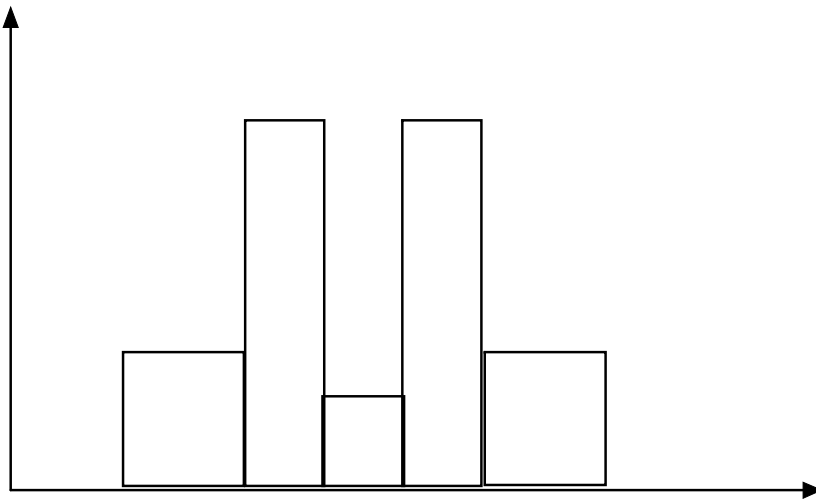
Histogram Analysis

- ▶ **Histogram**: summarizes the distribution of a given attribute
- ▶ Partition the data distribution into disjoint subsets, or buckets
- ▶ If the attribute is **nominal** → **bar chart**
- ▶ If the attribute is **numeric** → **histogram**



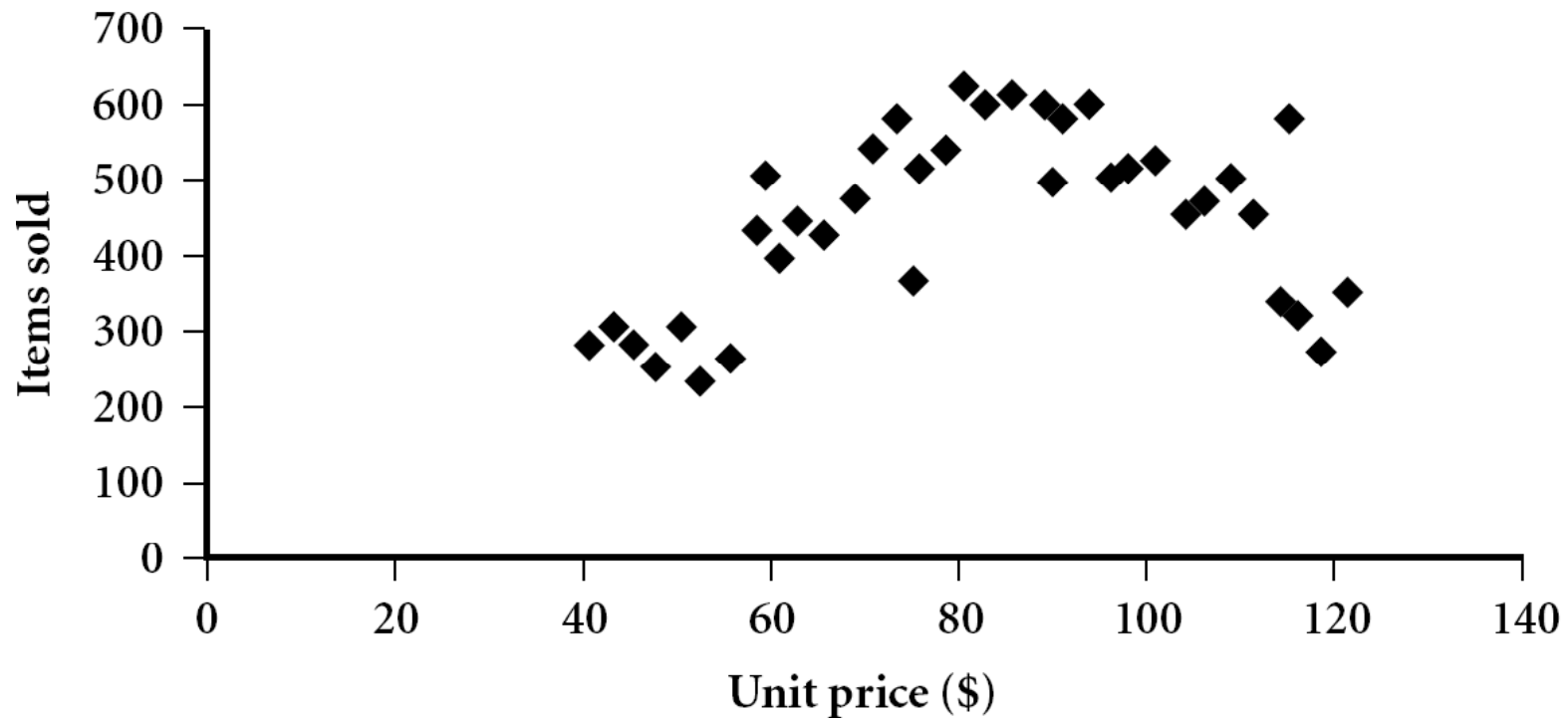
Histograms Often Tell More than Boxplots

- ▶ The two histograms shown in the left may have the same boxplot representation
 - The **same values** for: min, Q1, median, Q3, max, **But they have rather different data distributions**



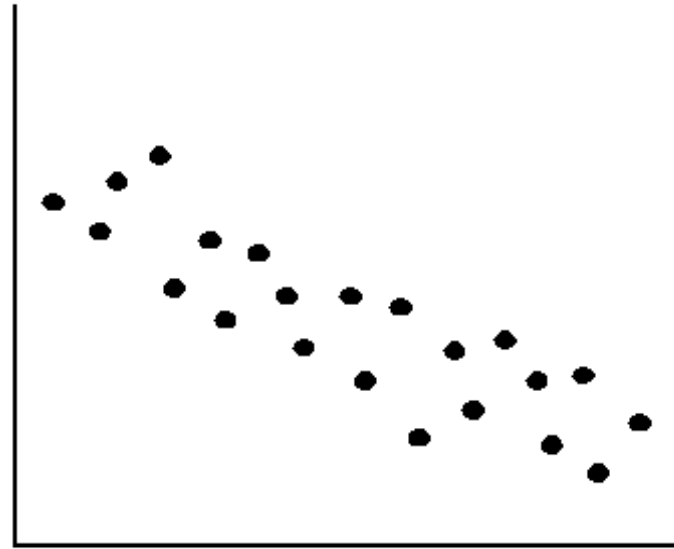
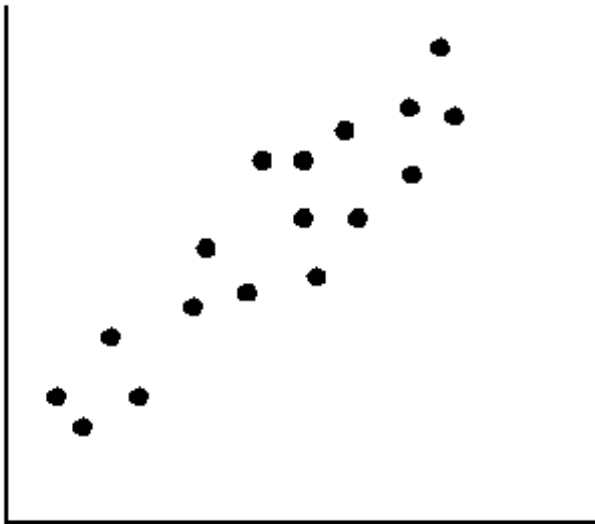
Scatter plot

- ▶ Provides a first look at bivariate data to see clusters of points, outliers, etc.
- ▶ Each pair of values is treated as a pair of coordinates and plotted as points in the plane

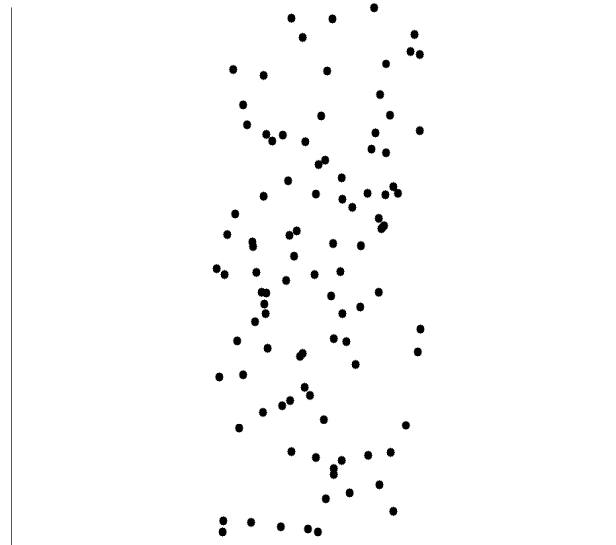
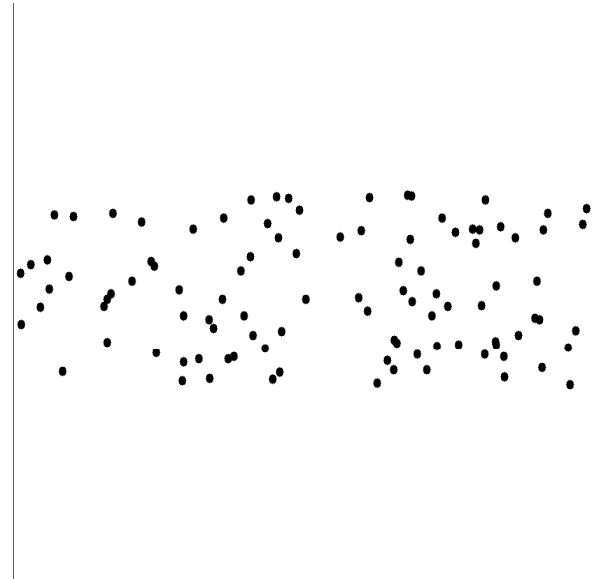
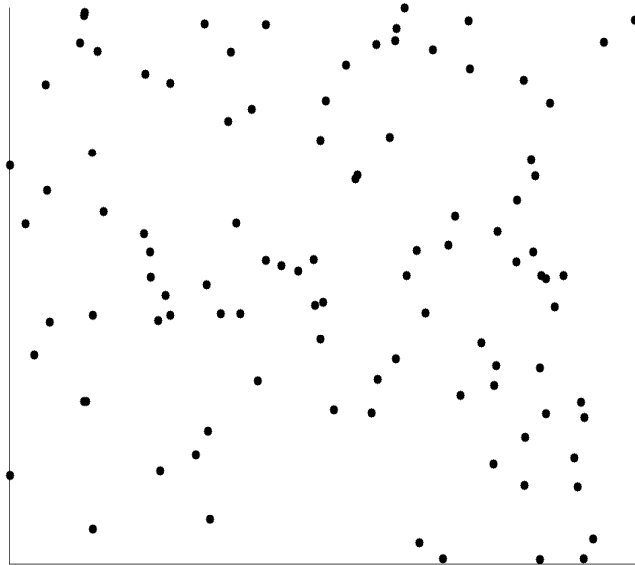


Positively & Negatively Correlated Data

- ▶ The left half fragment is positively correlated
- ▶ The right half fragment is negatively correlated



Uncorrelated Data



Summary of Section 1.2

- ▶ Data attribute types: nominal, binary, ordinal, interval-scaled, ratio-scaled
- ▶ Many types of data sets, e.g., numerical, text, graph, Web, image.
- ▶ Gain insight into the data by:
 - Measure data similarity
 - Basic statistical data description: central tendency, dispersion, graphical displays
 - Data visualization: map data onto graphical primitives
- ▶ Above steps are the beginning of data preprocessing.
- ▶ Many methods have been developed but still an active area of research.