Getting To Know Your Data

Road Map

1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity

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

Document data

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

record

Transactional data

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

record

Cross tabulation

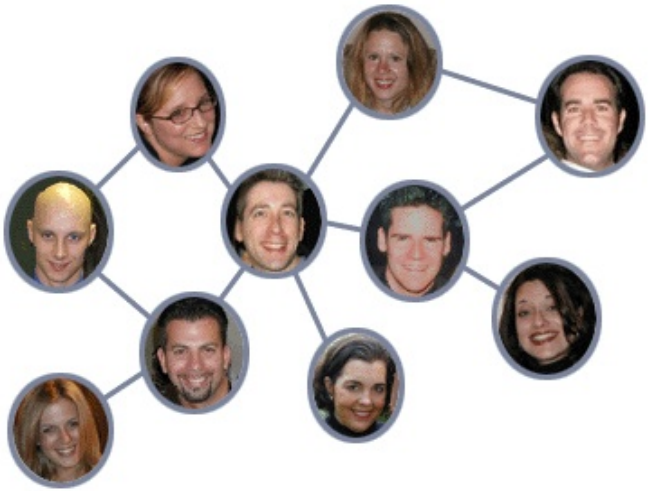
	Books	Multimedia devices
Big spenders	30%	70%
Budget spenders	60%	25%
Very Tight spenders	10%	5%

record

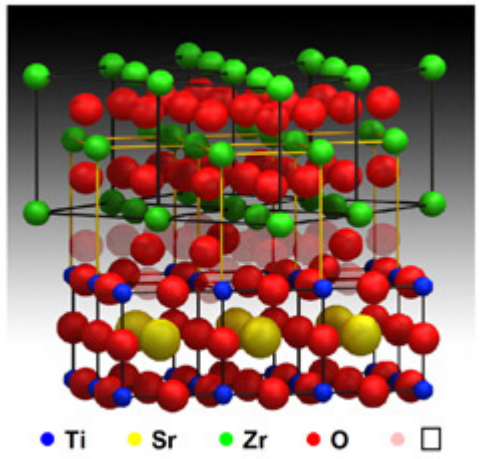
Types of Data Sets

- **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

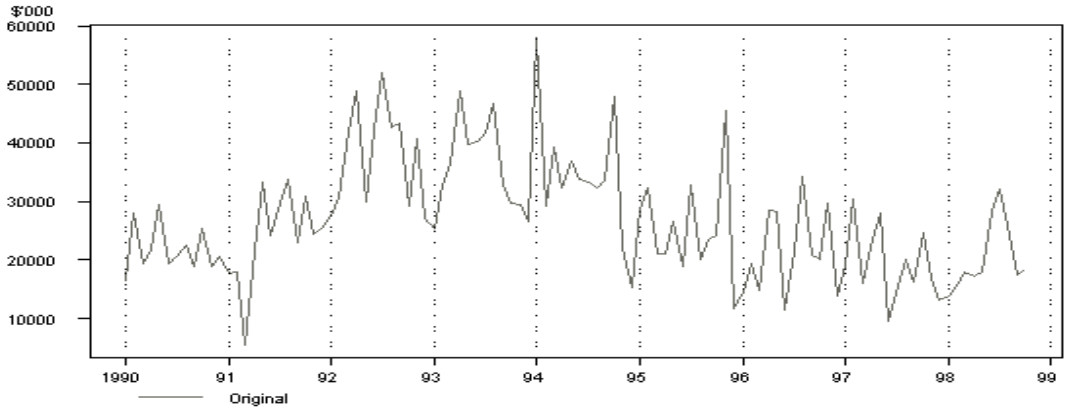
- **Ordered**
  - Videos
  - Temporal data
  - Sequential data
  - Genetic sequence data

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 and Attributes

- Datasets are made up of data objects.
- A **data object** (or **sample** , **example**, **instance**, **data point**, **tuple**) represents an entity.
- **Examples**
  - Sales database: customers, store items, sales
  - Medical database: patients, treatments
  - University database: students, professors, courses
- Data objects are described by **attributes** (or **dimension**, **feature**, **variable**).
- 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

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., having cancer)
- **Ordinal**
  - Values have a meaningful order (**ranking**) but magnitude between successive values is not known.
  - *Size = {small, medium, large}*, grades, army rankings

Attributes Types

- **Numeric:** quantity (integer or real-valued)

  Interval-Scaled

  - Measured on a scale of equal-sized units
  - Values have order
    - E.g., temperature in C˚or F˚, calendar dates
    - No true zero-point (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 **(10 K˚ is twice as high as 5 K˚)**
    - 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 countable 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)

Quiz

- What is the type of an attribute that describes the height of a person in centimeters?
  - Nominal
  - Ordinal
  - Interval-scaled
  - Ratio-scaled

- In Olympic games, three types of medals are awarded: bronze, silver, or gold. To describe these medals, which type of attributes should be used?
  - Nominal
  - Ordinal
  - Interval-scaled
  - Ratio-scaled

Road Map

1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity

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 properties**
  - Centrality: use measures such as the median
  - Variance: use measures such as the quantiles
- **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)**

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

  Note: *n* is sample size

  - A **distributive** measure can be computed by partitioning the data into smaller subsets (e.g., **sum**, and **count**)
  - 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 xi is weighted**

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

  - Weighted arithmetic mean

- **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 dataset
  - Holistic measures are much more expensive to compute than distributive measures
  - Can be estimated by interpolation (for grouped data):

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

    - 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

Example

Age	frequency
1-5	200
6-15	450
16-20	300
21-50	1500
51-80	700

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:

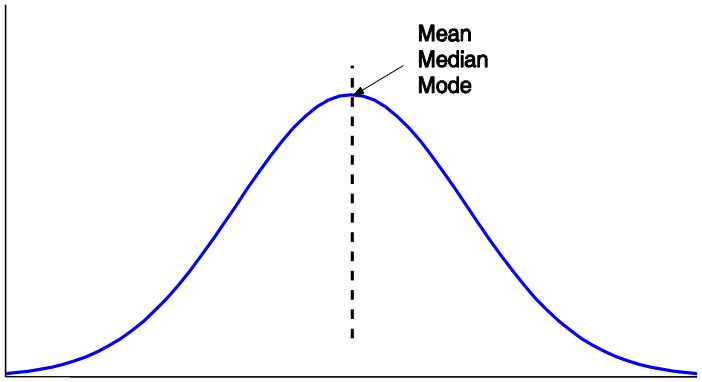
$$mean - mode = 3 \times (mean - median)$$

- **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 algebric measure that is easy to compute

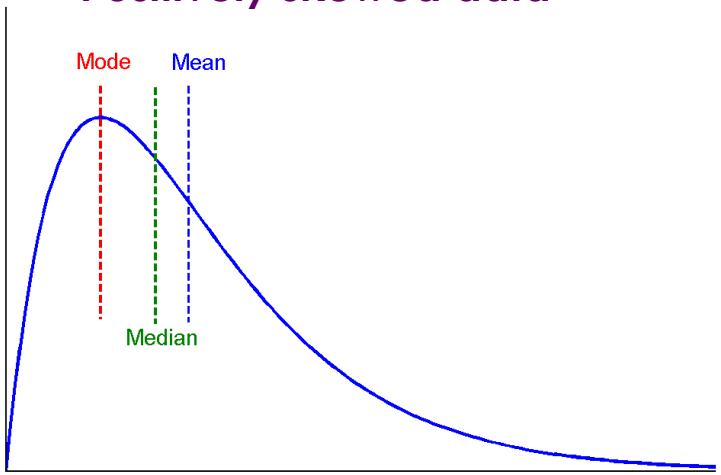
Symmetric vs. Skewed Data

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

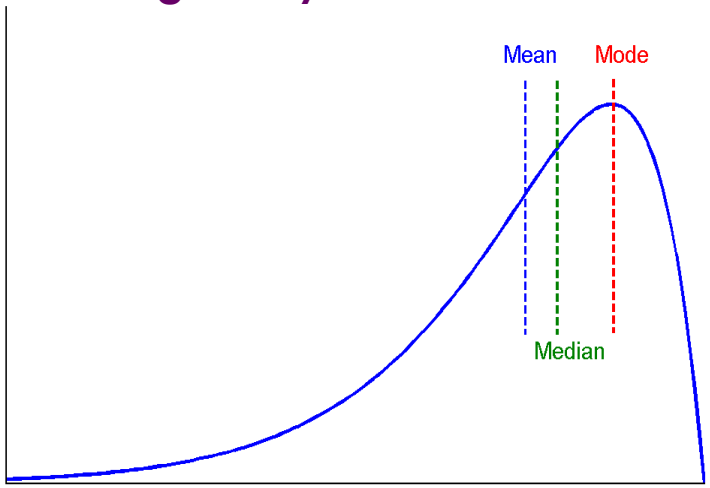
Symmetric data



Positively skewed data



Negatively skewed data



Quiz

- Give an example of something having a positively skewed distribution
  - income is a good example of a positively skewed variable -- there will be a few people with extremely high incomes, but most people will have incomes bunched together below the mean.
- Give an example of something having a bimodal distribution
  - bimodal distribution has some kind of underlying binary variable that will result in a separate mean for each value of this variable. One example can be human weight – the gender is binary and is a statistically significant indicator of how heavy a person is.

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
- **K^{th} 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 x 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., min= Q1-1.5 x IQR, max= Q3 + 1.5 x IQR
  - Represented by a Boxplot

Measuring the Dispersion of Data

- Variance and standard deviation
  - **Variance**: (algebraic, scalable computation)

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

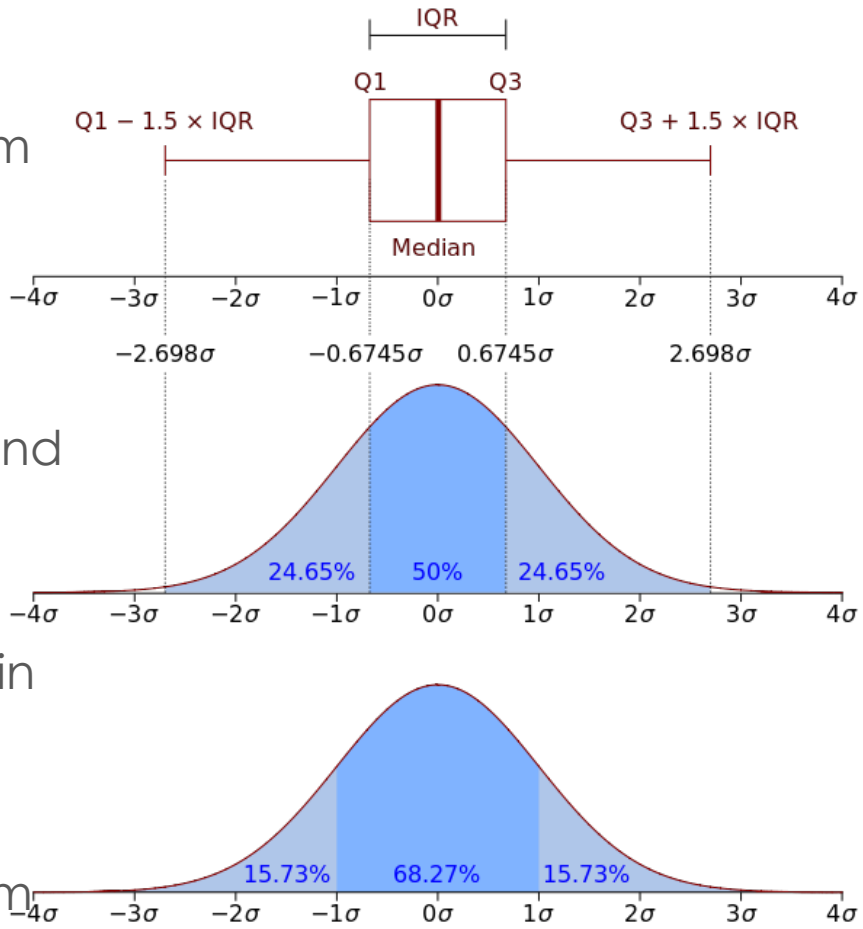
  - **Standard deviation** σ is the square root of variance σ^2
- **Basic properties of the standard deviation**
  - σ measures spread about the mean and should be used only when the mean is chosen as the measure of the center
  - $\sigma=0$ only when there is no spread, that is, when all observations 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

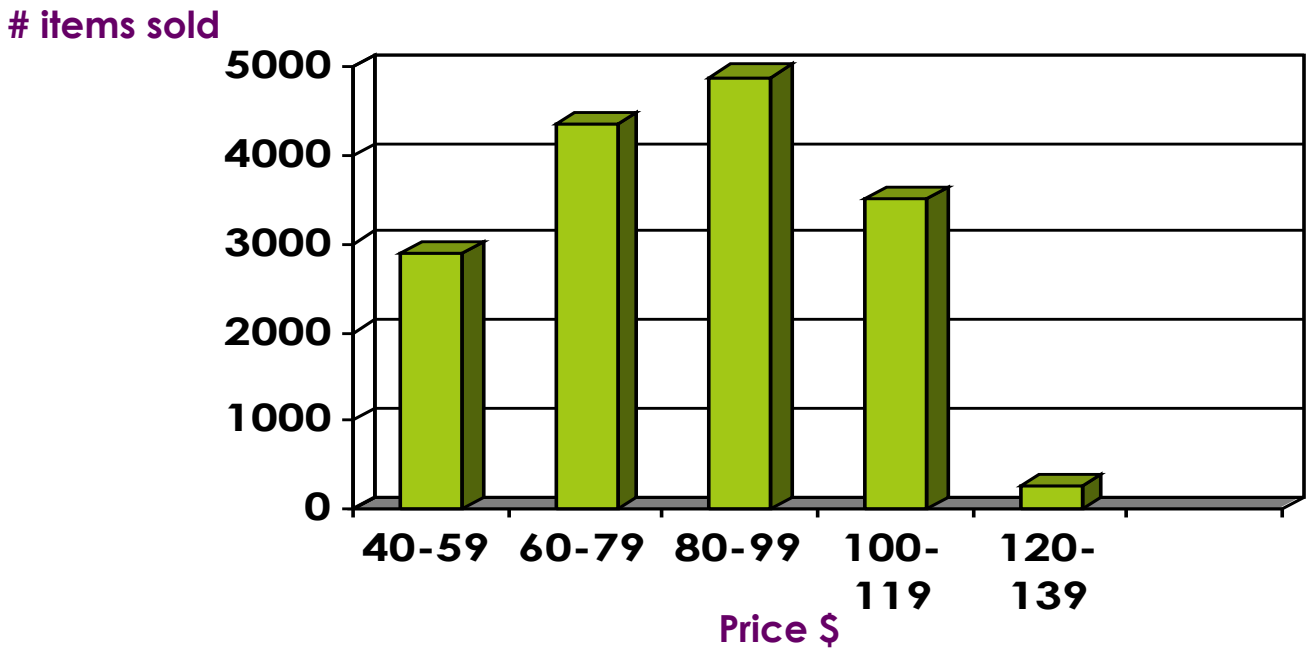
Boxplot

- **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



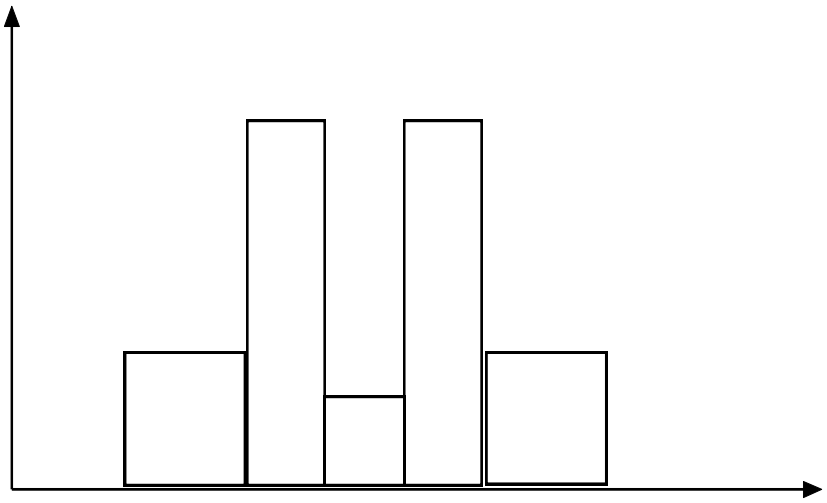
Histogram Analysis

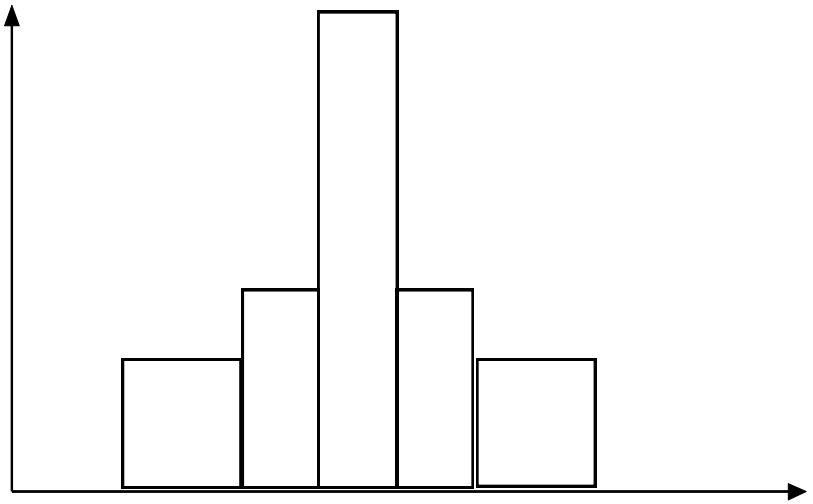
- **Histogram**: summarizes the distribution of a given attribute
- Partition the data distribution into disjoint subsets, or buckets
- If the attribute is **nominal** → **bar chart**
- If the attribute is **numeric** → **histogram**



Histograms Often Tell More than Boxplots

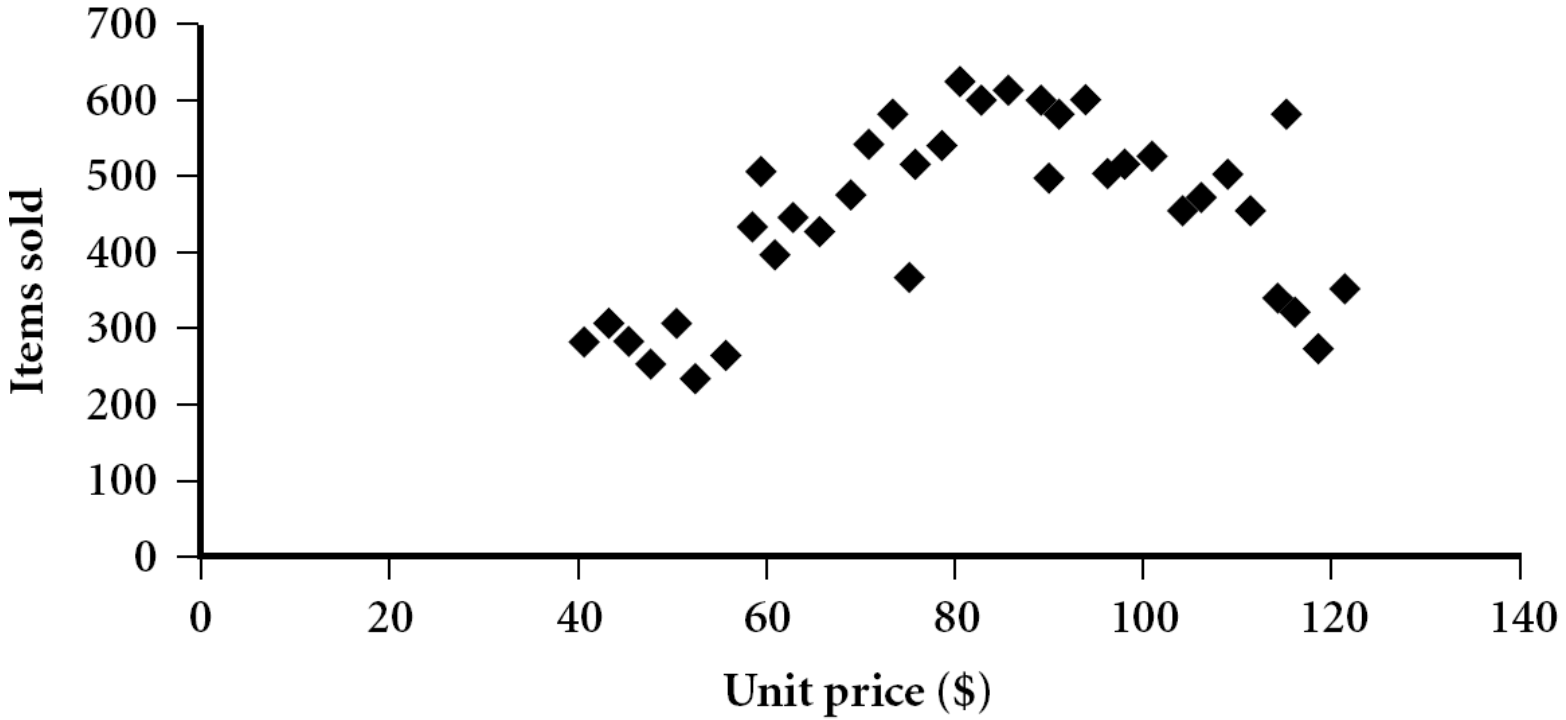
- The two histograms shown in the left may have the same boxplot representation
  - The **same values** for: min, Q1, median, Q3, max, **But they have rather different data distributions**





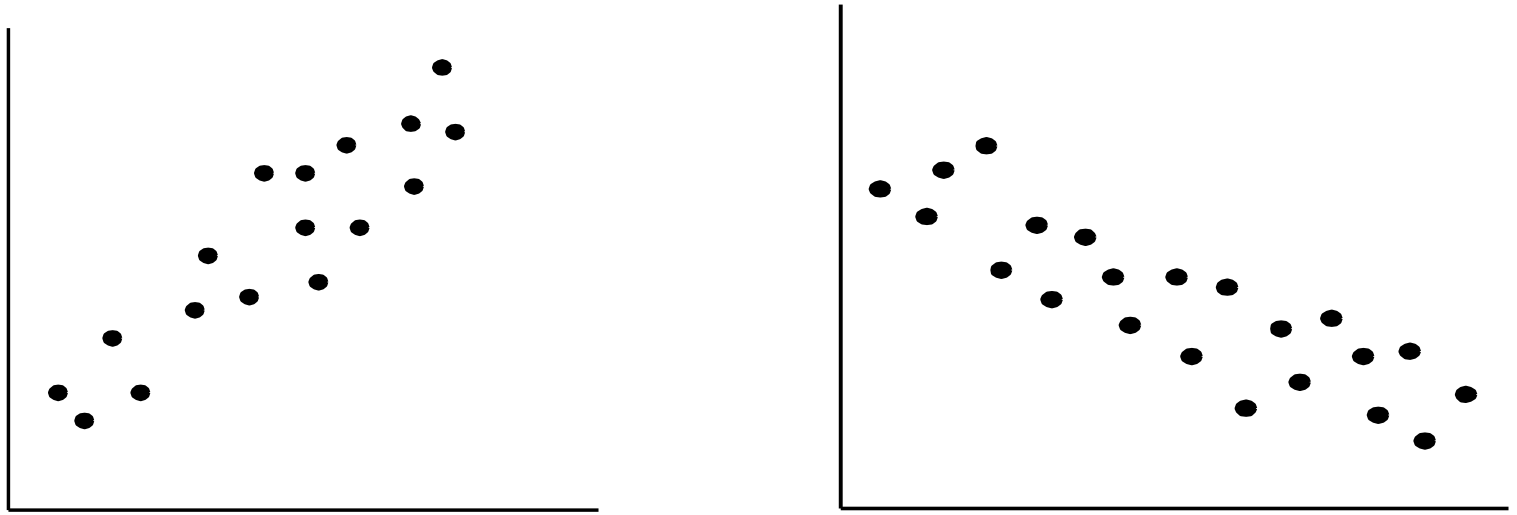
Scatter plot

- 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

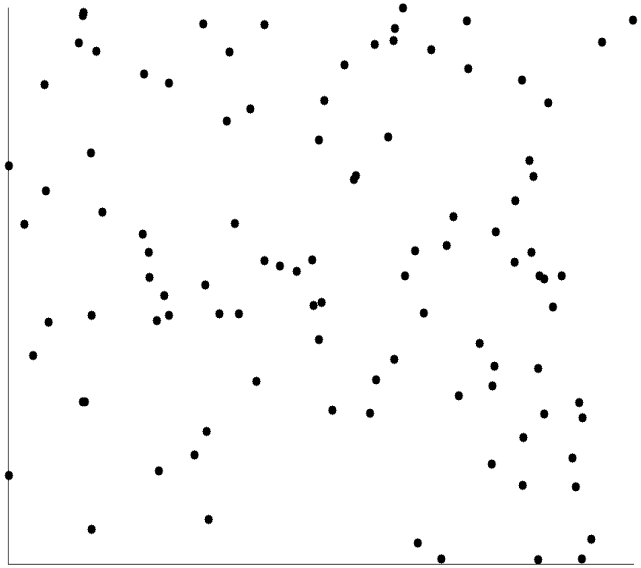


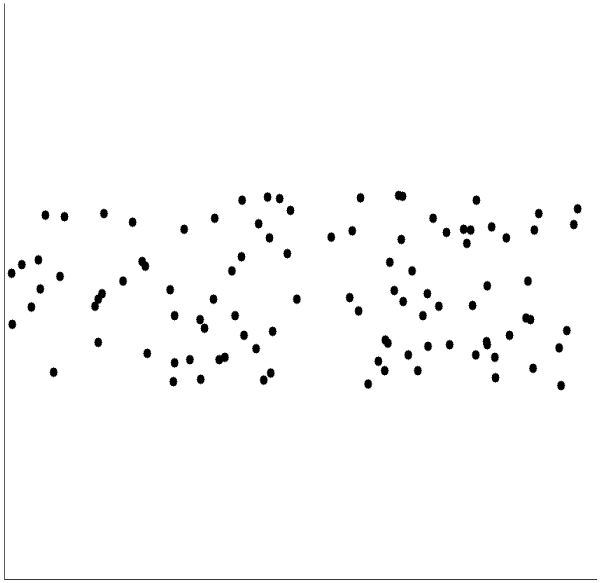
Positively & Negatively Correlated Data

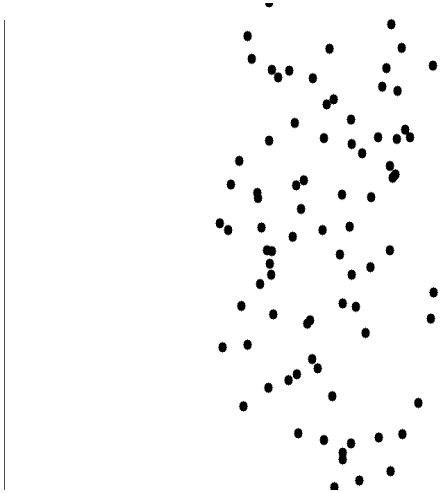
- The left half fragment is positively correlated
- The right half is negatively correlated



Uncorrelated Data







Road Map

1. Data Objects and Attribute Types
2. Descriptive Data Summarization
3. Measuring Data Similarity and Dissimilarity

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

Attributes

Data objects

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

- **Dissimilarity matrix**
  - n data points, but registers only the distance
  - A triangular matrix
  - Single mode: row and columns represent the same entity

Data objects

Data objects

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

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 *i* \ Object *j*	1	0	sum
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)

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

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

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

m_f the mean of f
$X_{1f},...x_{nf}$: 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

$$d(i,j) = \sqrt[h]{(|x_{i1} - x_{j1}|^h + |x_{i2} - x_{j2}|^h + \ldots + |x_{ip} - x_{jp}|^h)}$$

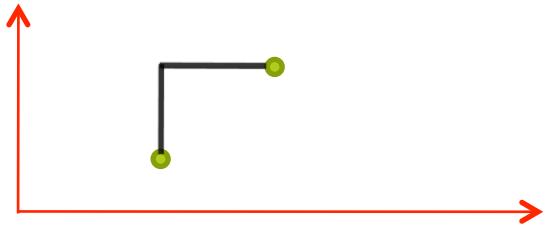
where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two p-dimensional data objects, and h is the order

- **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

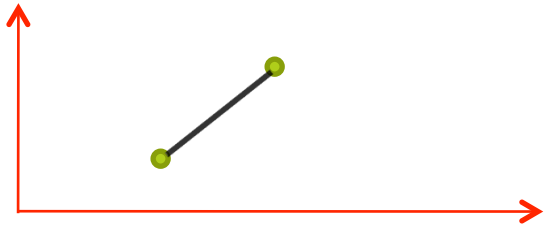
- ***h* = 1**: **Manhattan** (city block, L_1 norm) distance
  - E.g., the Hamming distance: the number of bits that are different between two binary vectors

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$



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

$$d(i,j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \ldots + |x_{ip} - x_{jp}|^2)}$$



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

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
  - map the range of each variable onto [0, 1] by replacing *i*-th object in the *f*-th variable by

$$r_{if} \in \{1, \ldots, M_f\}$$

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

  - compute the dissimilarity using methods for interval-scaled variables

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.

	team	coach	baseball	soccer	penalty	score	win	loss
Doc1	5	0	0	2	0	0	2	0
Doc2	3	0	2	1	0	0	3	0
Doc3	0	7	0	1	0	0	3	0
Doc4	0	1	0	1	2	2	0	3

- **Other vector objects**: gene features in micro-arrays, …
- **Cosine measure** : If d1 and d2 are two vectors (e.g., term-frequency vectors), then

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

where • indicates **vector dot product**, ||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 $d_1 \bullet d_2$**

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

- **Compute $||d_1||$**

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

- **Compute $||d_2||$**

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

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

Summary

- 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:
  - Basic statistical data description: central tendency, dispersion, graphical displays
  - Data visualization: map data onto graphical primitives
  - Measure data similarity
- Above steps are the beginning of data preprocessing.
- Many methods have been developed but still an active area of research.