Exploratory Data Analysis

Nguyen Ngoc Thao
nnthao@fit.hcmus.edu.vn

Content outline

- Data objects and Attributes
- Basic statistical data descriptions
- Basic data visualization
- Data proximity measures

Data collection: Record datasets

- Relational / transactional tuples
- Term-frequency vectors, numerical matrices, crosstabs

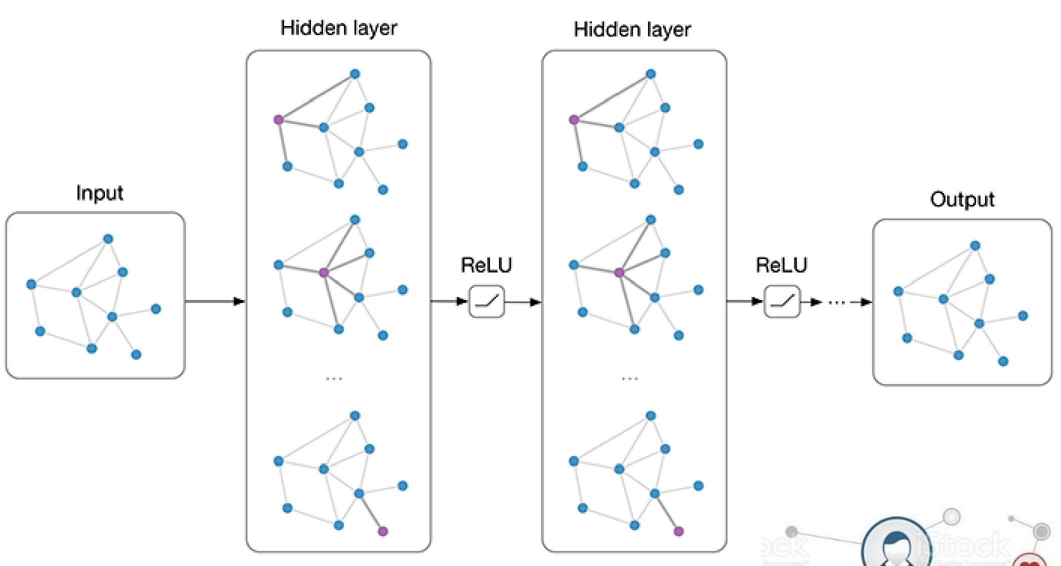
TID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}

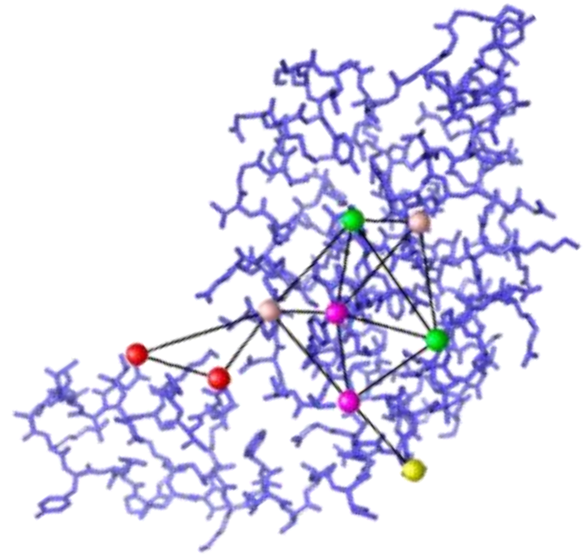
	for	great	greatest	lasagna	life	love
sentence 1	0	0	1	0	1	1
sentence 2	0	2	0	0	0	1
sentence 3	0	0	1	0	0	1
sentence 4	1	0	0	1	0	1

		Task Performance		**Total**
		Fail	**Success**	
User Felt	Very bad	0	0	0
	Bad	2	1	3
	Neutral	1	4	5
	Good	0	15	15
	Very good	0	5	5
Total		3	25	28

Data collection: Graph datasets

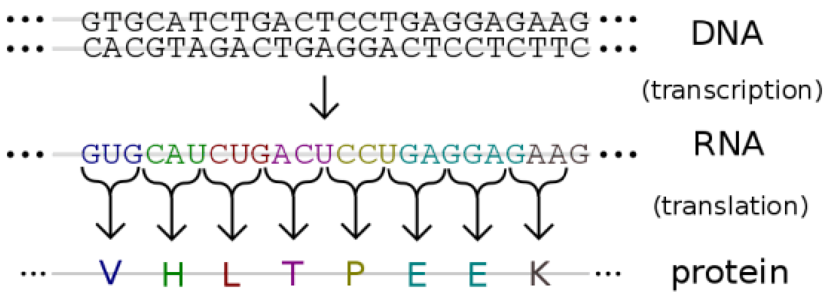
- The Internet, social networks, molecular structures

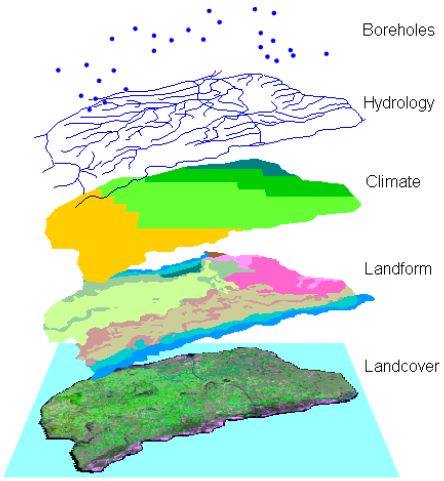


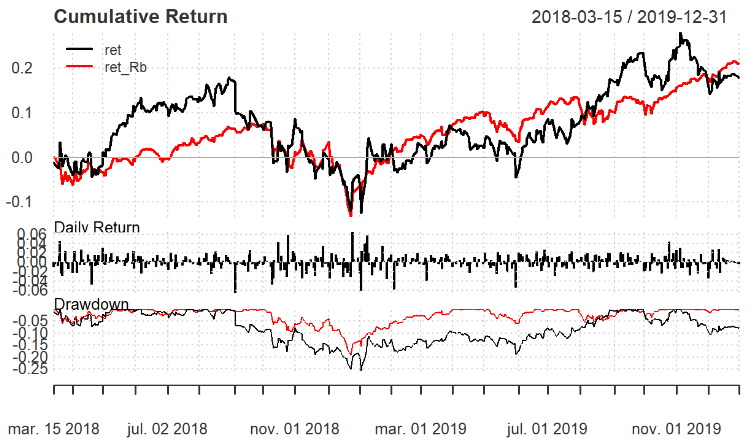


Data collection: Ordered datasets

- Sequential data: transaction sequences, genetic sequences
- Video data, temporal data, time-series data, etc.







Data objects and Attributes

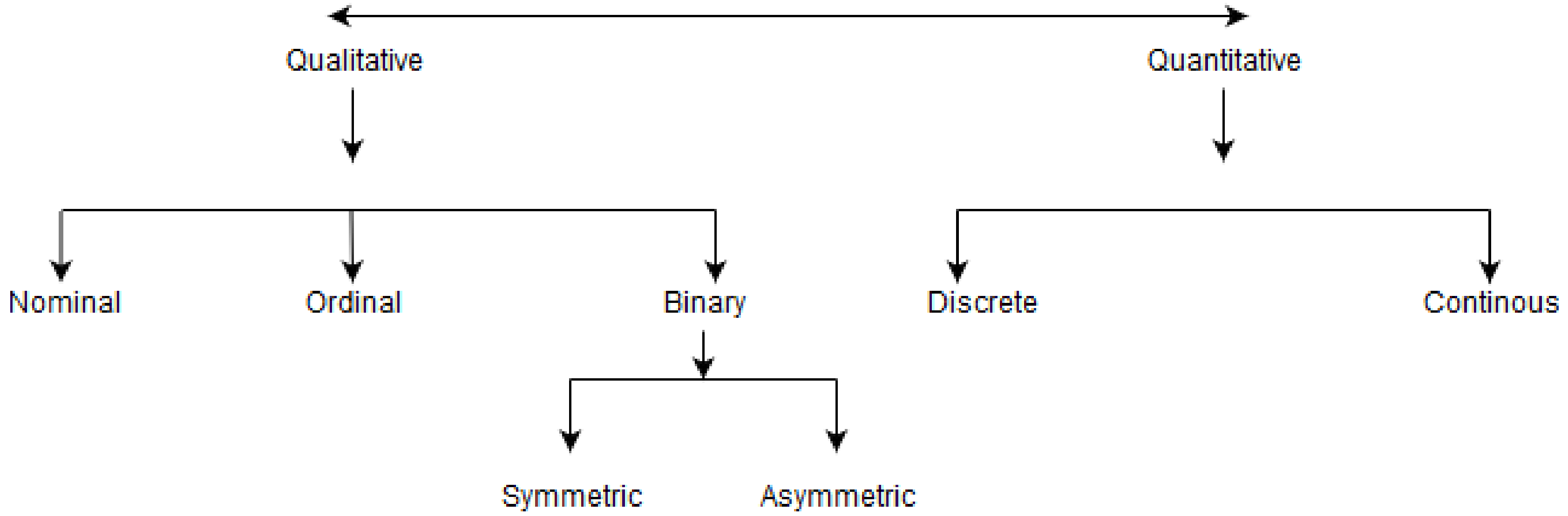
Data objects

- A data object represents an entity, serving as the building block for a dataset.
  - Similar terms: sample, example, instance, data point, and tuple

- Data objects are described by attributes.
  - In a database: rows → data objects, columns → attributes

Attributes

- An attribute shows some characteristic of a data object.
  - Similar terms: dimension, feature, and variable
  - E.g., a Customer object has 3 attributes {id, name, address}
- Observation: an observed value for a given attribute
- Feature vector: a set of attributes used to describe an object

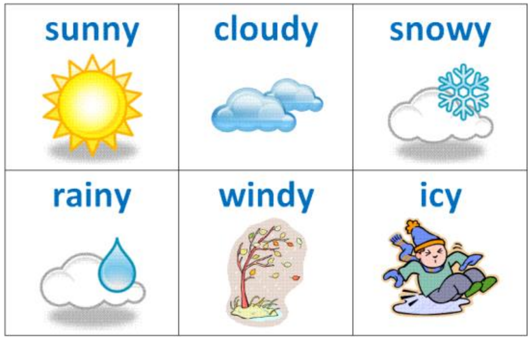


Attribute types: Nominal

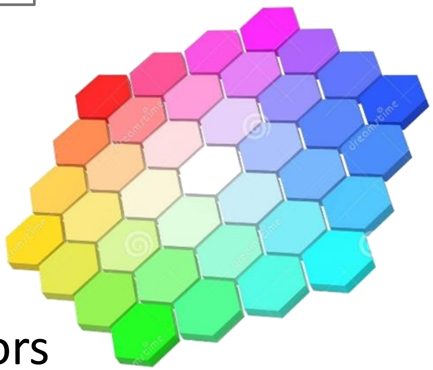
- Qualitative, values do not have any meaningful order
- Enumerations: categories, states, or “names of things”

Day and Night

Occupation



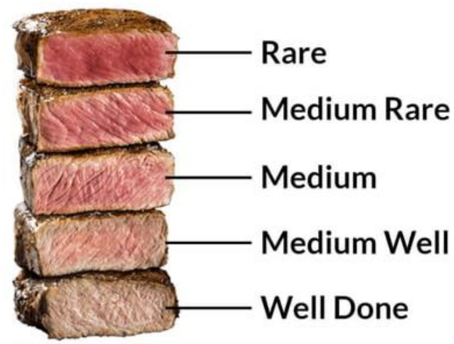
Weather

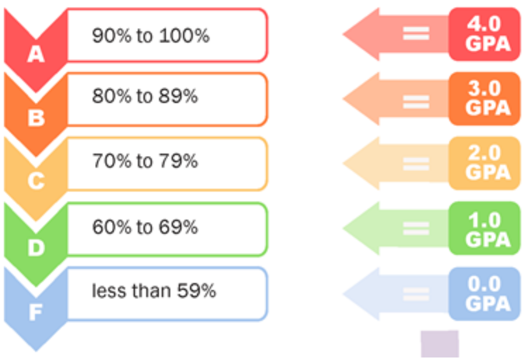


Colors

Attribute types: Ordinal

- Qualitative, values have a meaningful order (ranking) but magnitude between successive values is not known



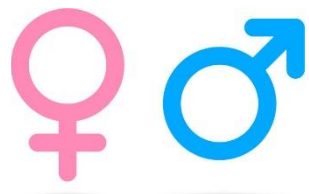


- Useful for subjective assessments of qualities that cannot be measured objectively
  - E.g., customer satisfaction

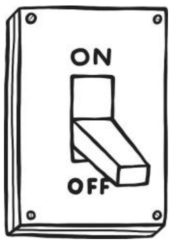
Attribute types: Binary

- Nominal attribute with only 2 states
- Symmetric binary: both outcomes equally important

Day and night



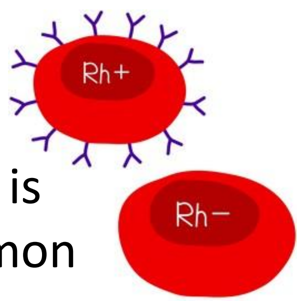
Male and Female

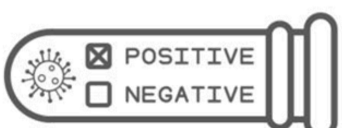


Switch light
On and Off

- Asymmetric binary: outcomes not equally important
  - Convention: assign 1 to the most important outcome (e.g., HIV test)

Rh positive is
more common





A positive result is more significant

Attribute types: Numeric

Interval numeric attribute

- Measured on a scale of **equal-sized units**
- Values have order (e.g., temperature in C˚ or F˚, calendar dates)
- No true **zero-point**: able to compute the difference – not able to talk of one value as being a multiple of another
  - E.g., 20˚C is five degrees higher than 15˚C (right), 10˚C is twice as warm as 5˚C (wrong)

Ratio numeric attribute

- Inherent **zero-point**
- Values can be considered as being an order of magnitude larger than the unit of measurement
  - E.g., temperature (10˚K is twice as high as 5˚K), monetary (you are 100 times richer with $100 than with $1), measurements (height, weight)

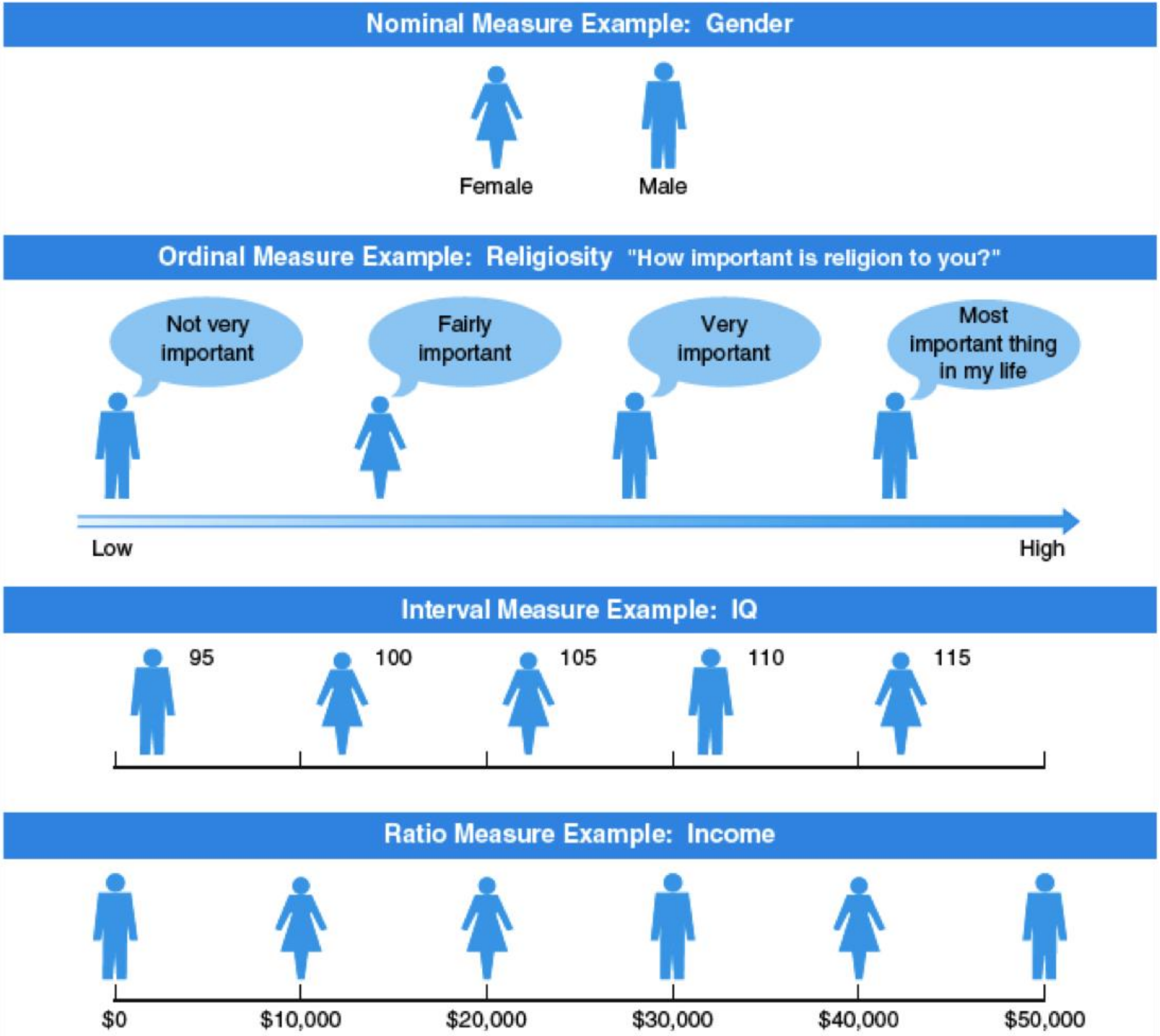


Image credit: Google Sites

Attributes: Discrete vs. Continuous

- There are many ways to organize attribute types, which are not mutually exclusive.
- Discrete attribute
  - Only a finite or countably infinite set of values
  - The values are sometimes represented as integers.
  - Binary attributes are a special case of discrete attributes.
- Continuous attribute
  - Real numbers of continuous domains
  - The values are usually represented using a finite number of digits

    $\rightarrow$ floating-point variables

Basic statistical data descriptions

Central tendency: Arithmetic mean

- Let $x_1, x_2, \dots, x_N$ be a set of N values or observations for some numeric attribute X.
- The arithmetic mean is defined as $\boldsymbol{\mu = \frac{1}{N}\sum_{i=1}^{N} x_i}$
- The weighted arithmetic mean is written as $\mu^w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - where w_i is the weight value that associates with x_i.
- It is the most common and effective numeric measure

Central tendency: Arithmetic mean

- Consider the score records of John and Kelly.
- The (non-weighted) mean scores are

$$\mu_{John} = 82.6, \qquad \mu_{Kelly} = 84.6$$

John's record	
Homework	92
Quiz	74
Lab	83
Test	76
Final exam	88

Kelly's record	
Homework	100
Quiz	82
Lab	95
Test	70
Final exam	76

Homework	15 %
Quiz	10 %
Lab	20 %
Test	25 %
Final exam	30 %

- We now have the course grade distribution
- The weighted mean scores are

$$\mu^w_{John} = 83.2, \qquad \mu^w_{Kelly} = 82.5$$

$$\mu^w_{John} = \frac{0.15 \times 92 + 0.1 \times 74 + 0.2 \times 83 + 0.25 \times 76 + 0.3 \times 88}{0.15 + 0.1 + 0.2 + 0.25 + 0.3} = 83.2$$

Central tendency: Arithmetic mean

- Means are highly sensitive to extreme values (e.g., outlier).
- Trimmed mean: chop extreme values before calculating the regular mean

Typical mean: 27.9

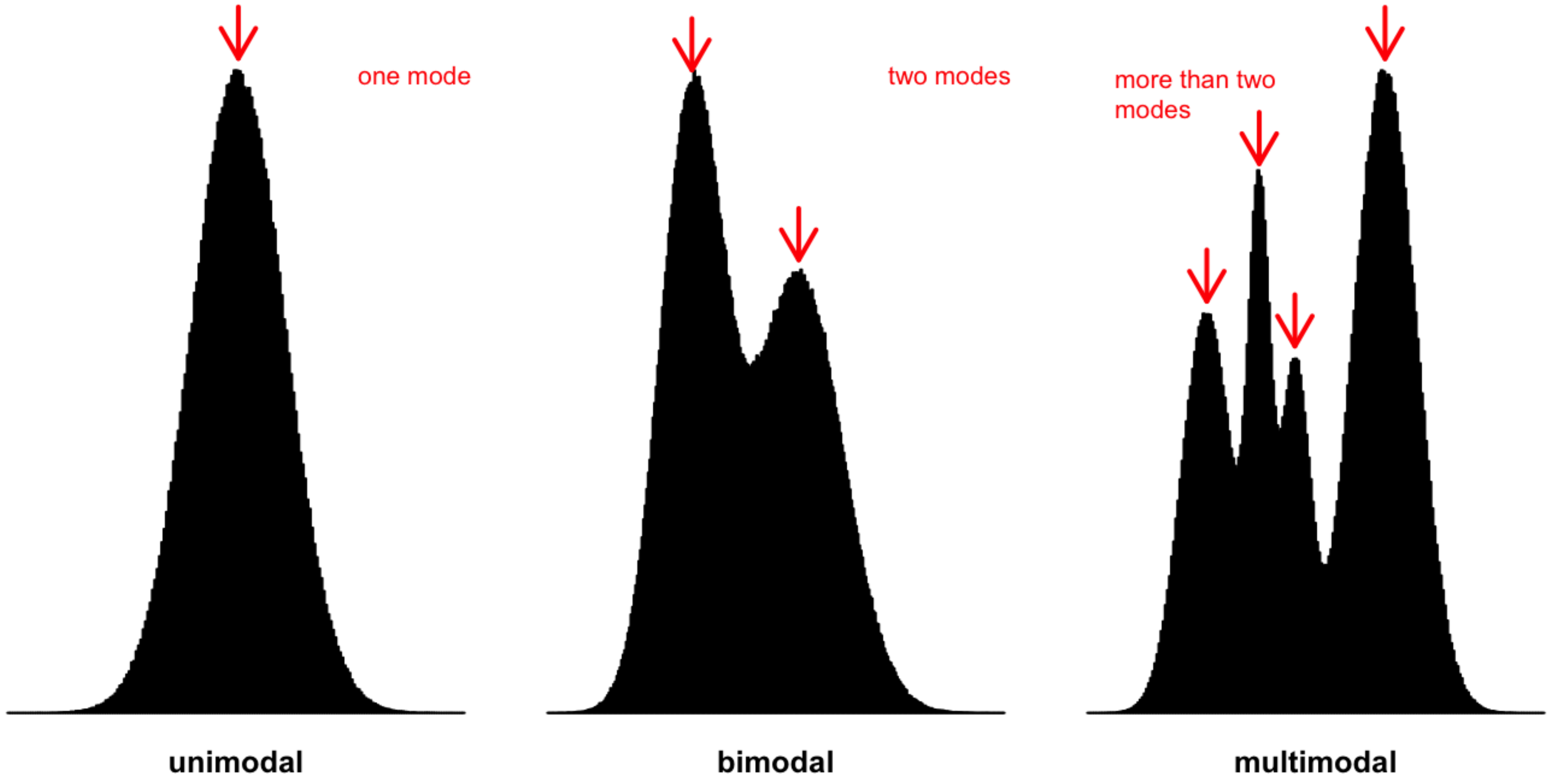
4 | 14 19 20 22 24 25 26 26 | 99

remove 10% observations from each side

Trimmed mean: 22

Central tendency: Mode

- Mode is the value that occurs most frequently in the data, defined for both qualitative and quantitative attributes.
  - If each data value occurs only once, then there is no mode



Central tendency: Median

- Suppose that the given set of N observations is sorted.
- Median is the middle value of the ordered set.
  - N is **odd**: pick the *exact middle value*; otherwise, take the *average of the two middlemost values*.
- Midrange is the average of the largest and smallest values in the set.

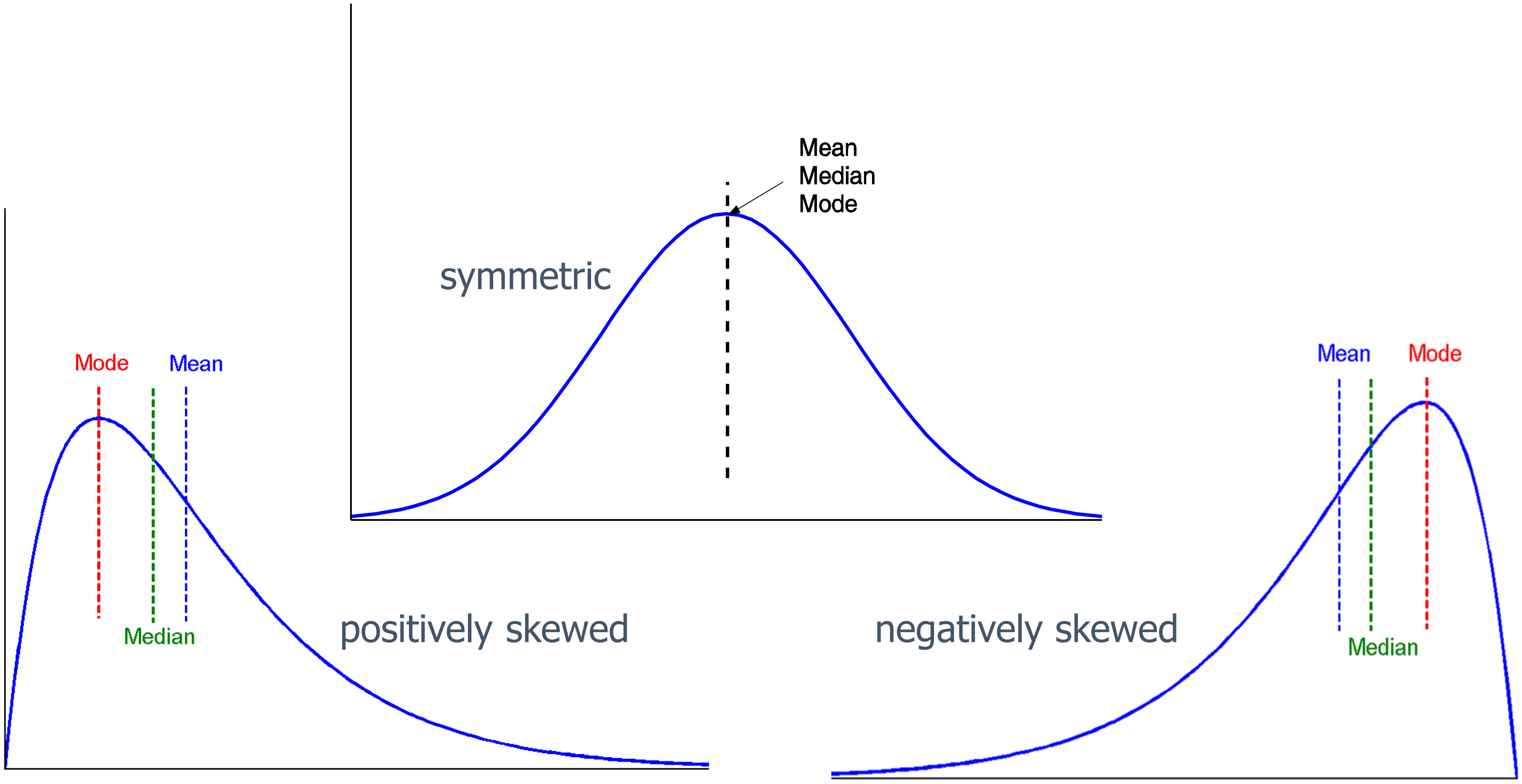
4 4 4 9 15 15 15 27 37 48

mean = 17.8 – mode: 4 and 15 – midrange = 22, median = (15+15)/2 = 15

3 3 6 9 15 15 15 27 27 37 48

mean = 18.636 – mode: 15 – midrange = 22.5, median = 15

Symmetric data vs. Skew data



- For moderately skewed unimodal numeric data, the empirical formula is

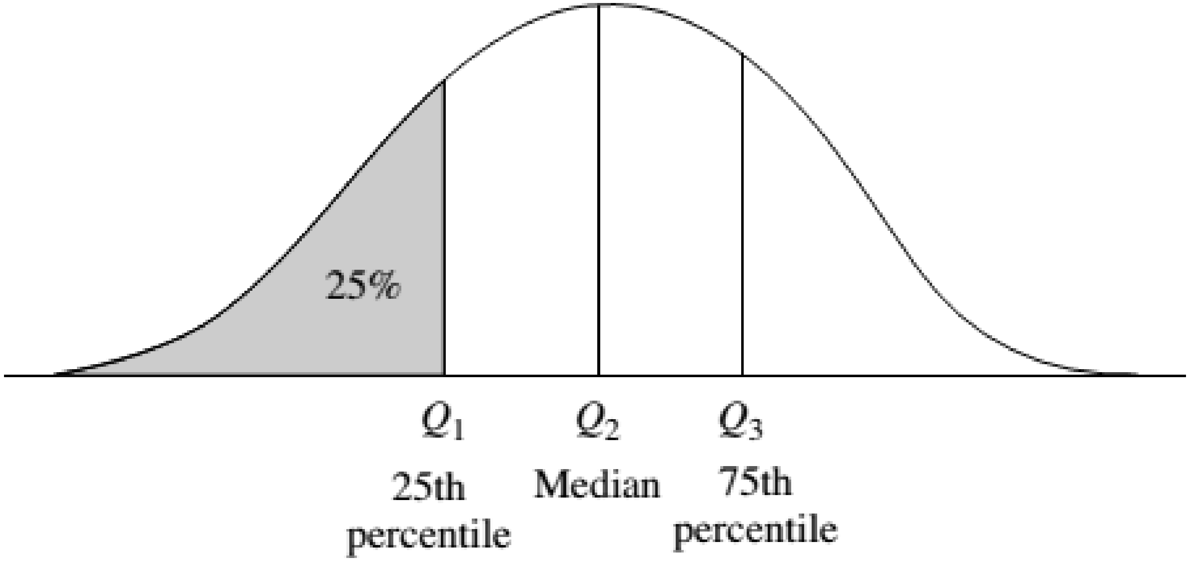
$$mean - mode \approx 3 \times (mean - median)$$

Data dispersion: Quantiles

- Let $x_1, x_2, \dots, x_N$ be a set of N observations sorted in increasing order for a numeric attribute X.
- Quantiles are points taken at regular intervals of a data distribution, dividing it into equal-sized consecutive sets.
- *k*th q-quantile ($0 < \mathrm{k} < q$, $k \in \mathbb{N}^*$): a value x such that at most k/q data values $< x$ and at most $(q - k)/q$ of which $> x$.
  - There are $q - 1$ q-quantiles.

Data dispersion: Quantiles

- Quartiles (4-quantiles) split the data distribution into four equal parts.



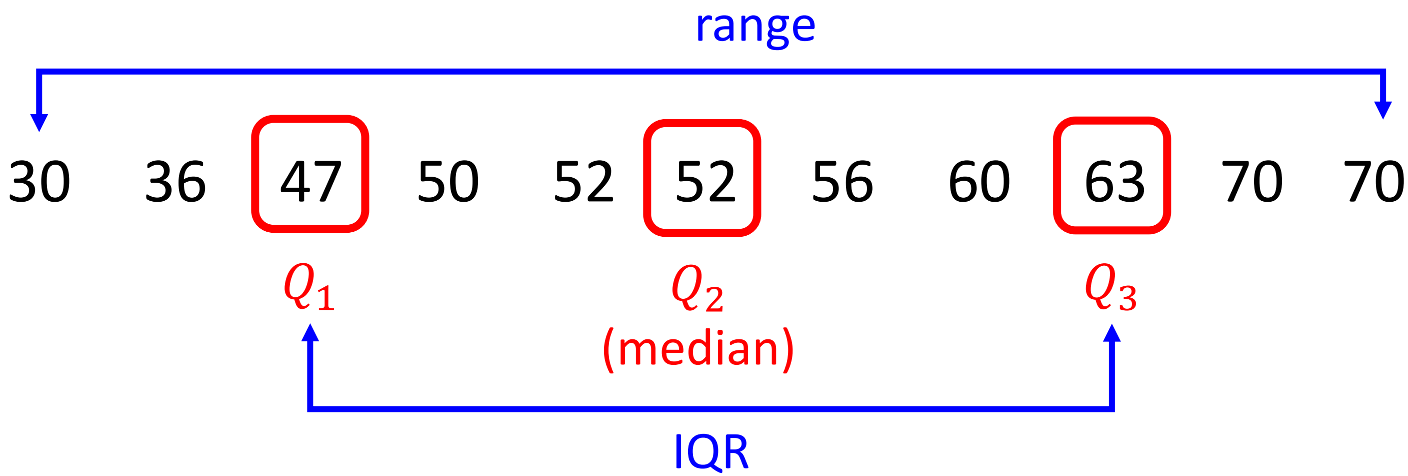
- Percentiles (100-quantiles): 100 equal-sized consecutive sets
- 2-quantile is the median that splits the distribution into halves

Data dispersion: Interquartile range

- Interquartile range (IQR) is the distance between the first and third quartiles.

$$IQR = Q_3 - Q_1$$

- Range is the difference between the largest and smallest values in the set.



How to determine the quartile?

- Use the median to divide the ordered set into two halves.
  - If the original set has an even number of points, split it exactly in half
  - Otherwise, **do not include** the median in either half.
- Q_1 and Q_3 are the medians of the lower and upper halves, respectively.

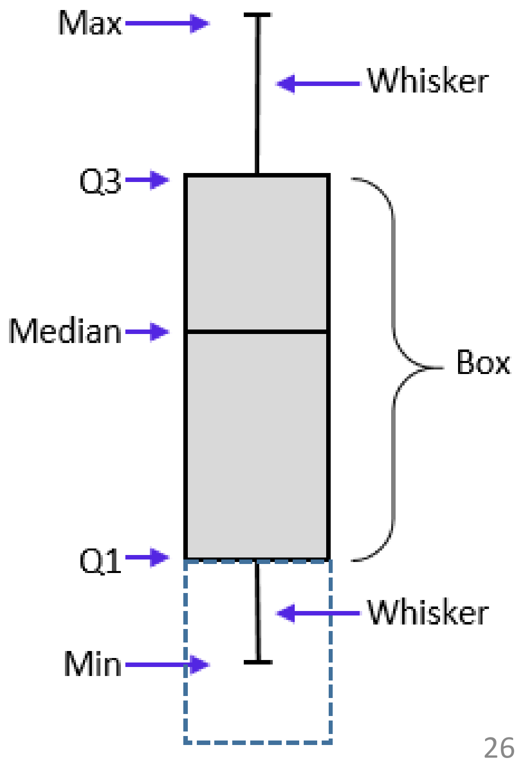
$$6 \quad 7 \quad \underset{Q_1}{\boxed{15}} \quad 36 \quad 39 \quad \underset{Q_2}{\boxed{40}} \quad 41 \quad 42 \quad \underset{Q_3}{\boxed{43}} \quad 47 \quad 49$$

$$7 \quad \underset{Q_1}{\boxed{15}} \quad \underset{Q_2 = 37.5}{\boxed{36 \quad 39}} \quad \underset{Q_3}{\boxed{40}} \quad 41$$

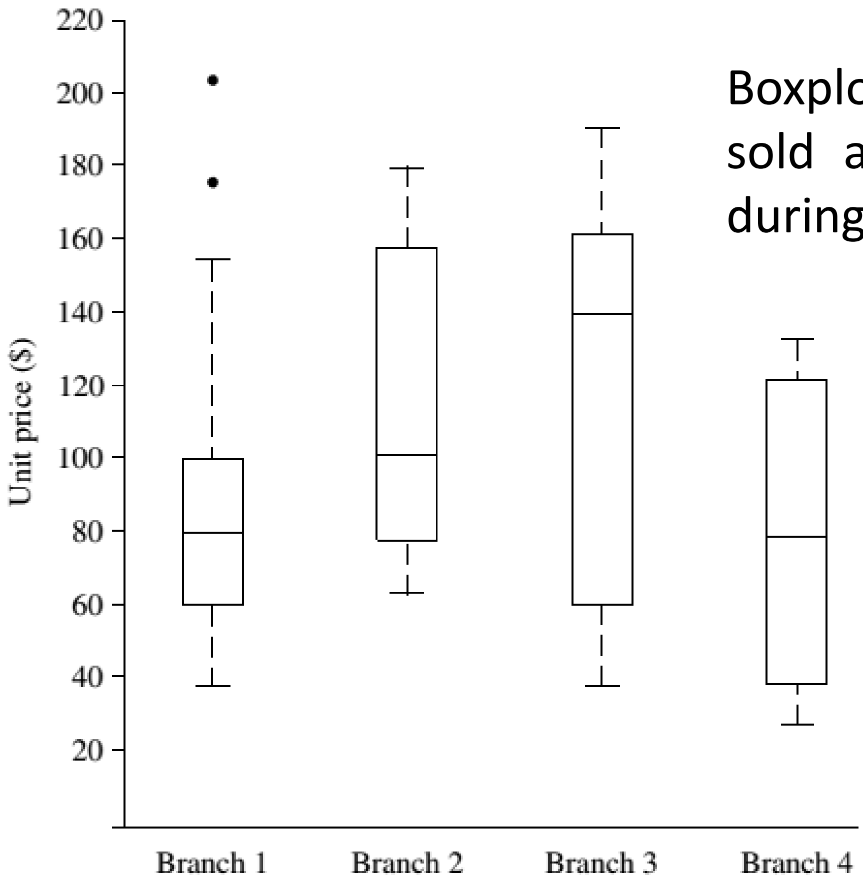
Data dispersion: Boxplot

- A five-number summary of a distribution includes
  - The median ($\boldsymbol{Q_2}$), the quartiles $\boldsymbol{Q_1}$ and $\boldsymbol{Q_3}$,
  - The smallest ($\boldsymbol{Min}$) and largest ($\boldsymbol{Max}$) individual values.

- The summary is presented by a boxplot.
  - Outliers: points that are out the range $[-1.5 \times IQR, 1.5 \times IQR]$, plotted individually



Data dispersion: Boxplot



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

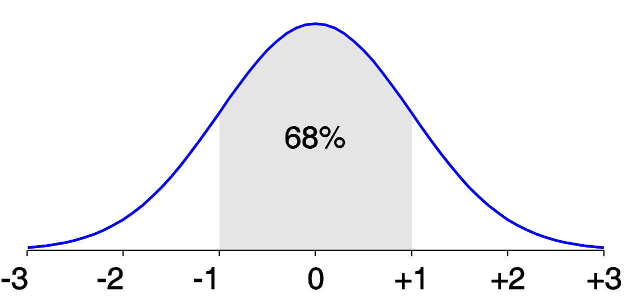
- For Branch 1, the median price of items sold is \$80, Q_1 is \$60, and Q_3 is \$100. Notice that two outlying observations, 175 and 202, were plotted individually as they are more than 1.5 × IQR.

Data dispersion: Variance

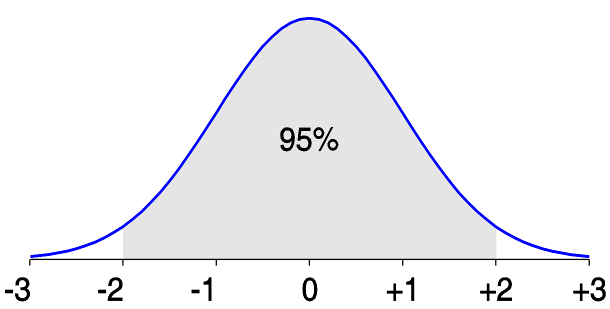
- The (population) variance is defined as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N} {x_i}^2\right) - \bar{x}^2$$

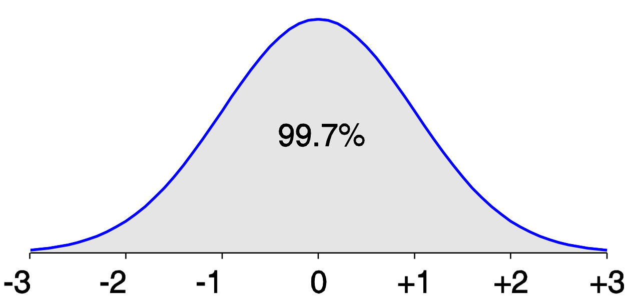
- The standard deviation is the square root of the variance.
  - *Low* $\sigma \rightarrow$ the data tends to be very *close to the mean*
  - *High* $\sigma \rightarrow$ the data *spreads out over a large range* of values



$[\mu - \sigma,\ \mu + \sigma]$



$[\mu - 2\sigma,\ \mu + 2\sigma]$



$[\mu - 3\sigma,\ \mu + 3\sigma]$

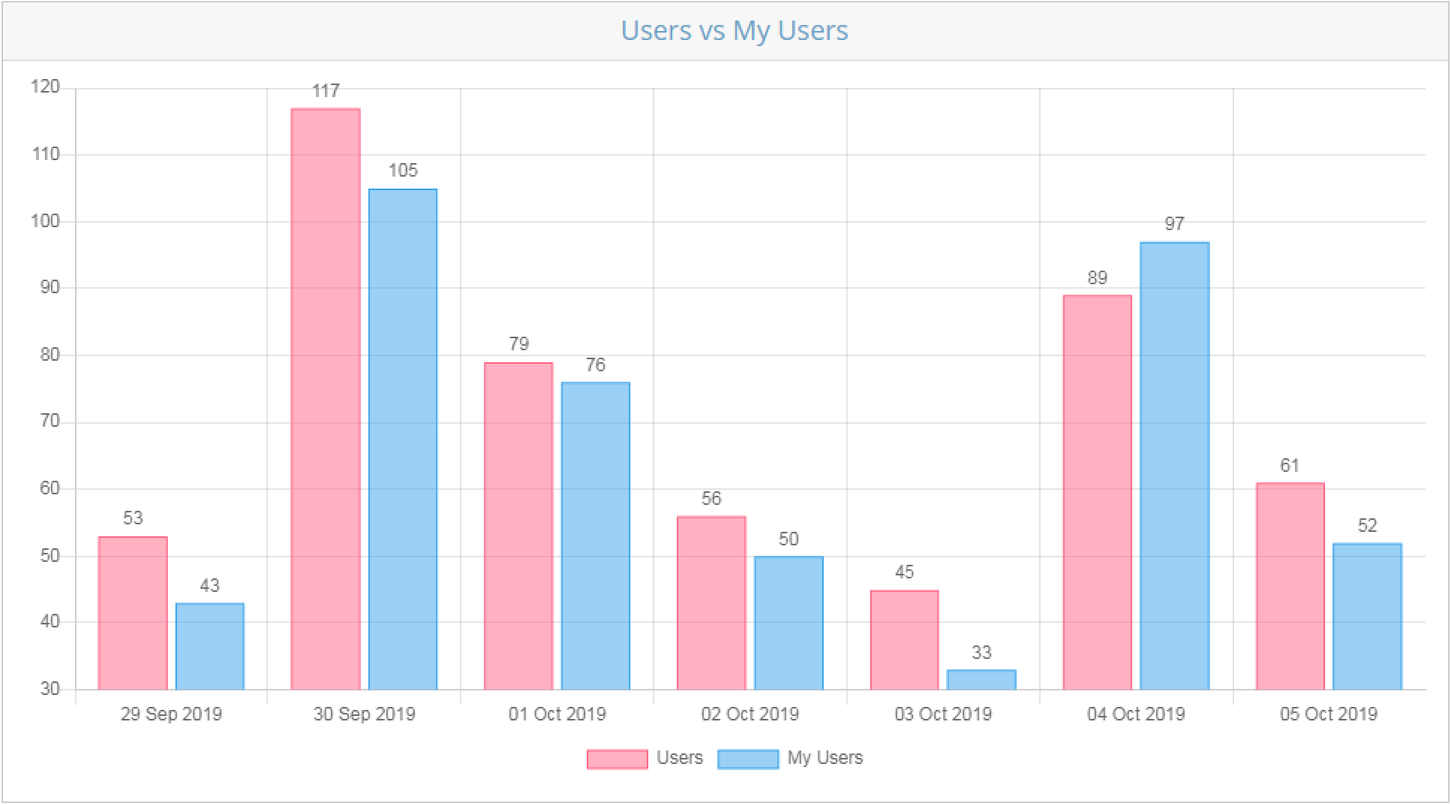
Basic data visualization

Why data visualization?

- Gain insight into an information space by mapping data onto graphical primitives
- Provide qualitative overview of large datasets
- Search for patterns, trends, irregularities, relationships among data
- Help find interesting regions and suitable parameters for further quantitative analysis
- Provide a visual proof of computer representations derived

Bar chart

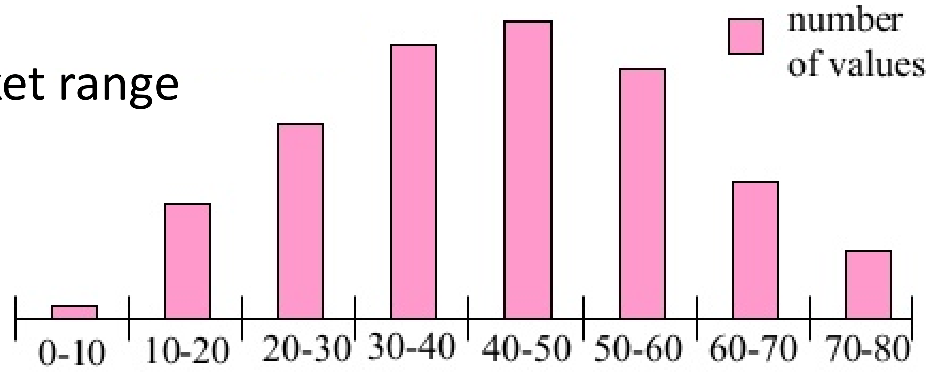
- A bar chart presents **nominal data** by using rectangular bars with heights proportional to the values represented.



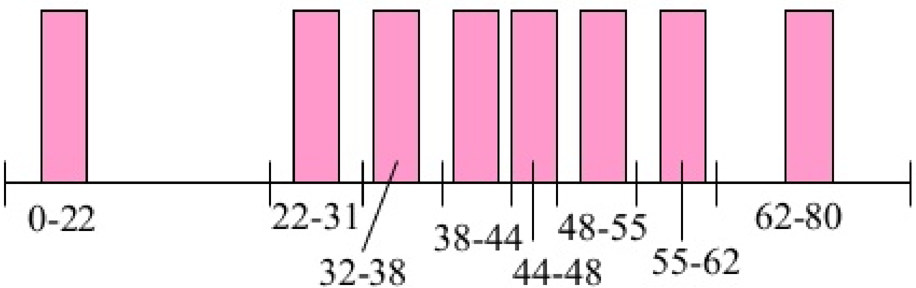
Histogram

- The range of values for a numeric attribute X is partitioned into disjoint consecutive subranges, called **buckets** or **bins**.
- A bar is drawn for each subrange such that its height represents the total items within the subrange.

Equal-width: equal bucket range



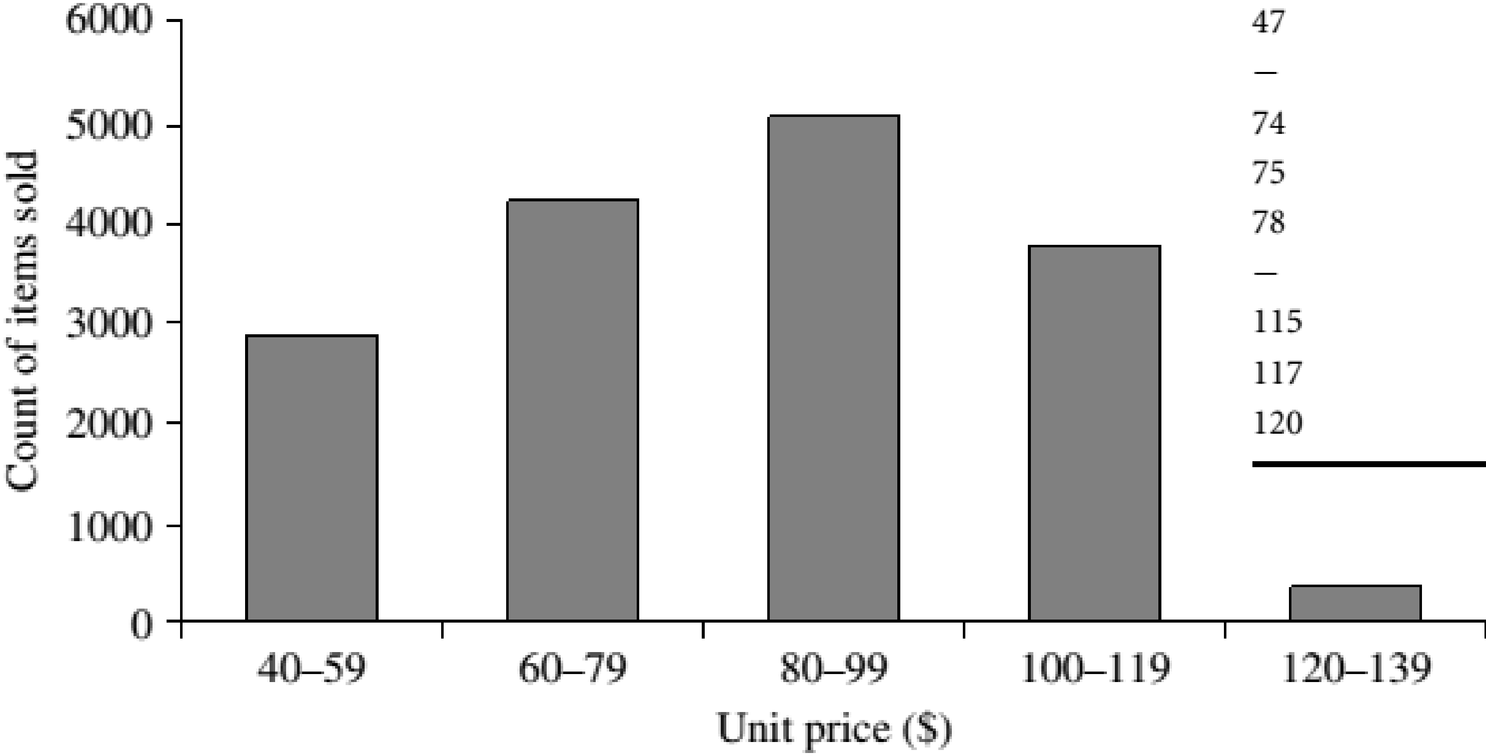
Equal-frequency: equal bucket depth



Histogram: An example

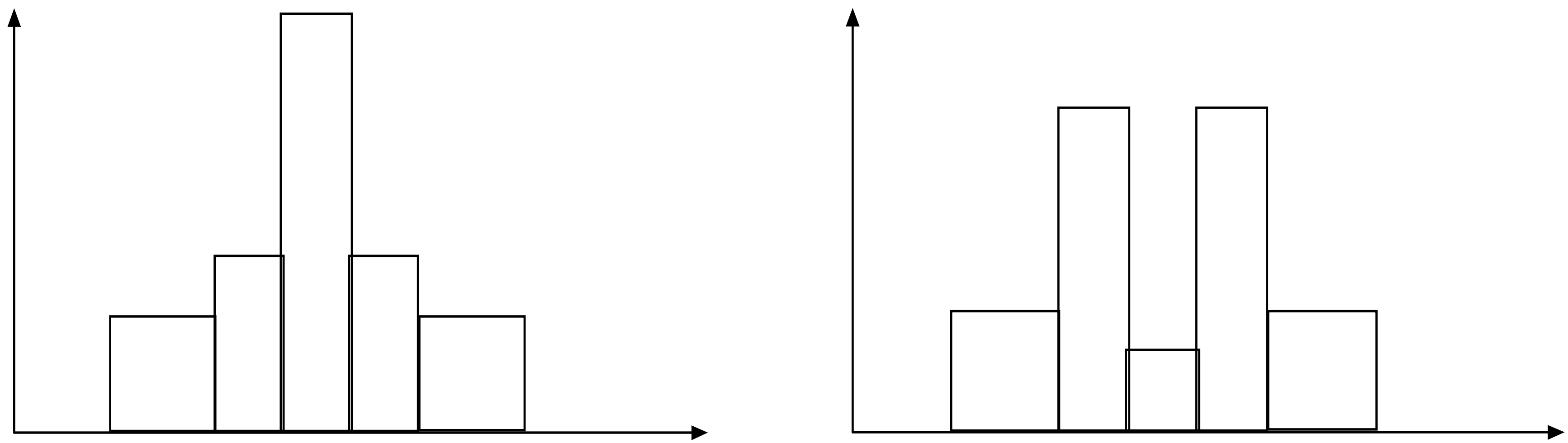
A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



Histogram over boxplot

- The two following histograms may have the same boxplot.
- However, they represent rather different data distributions.



Quantile plot

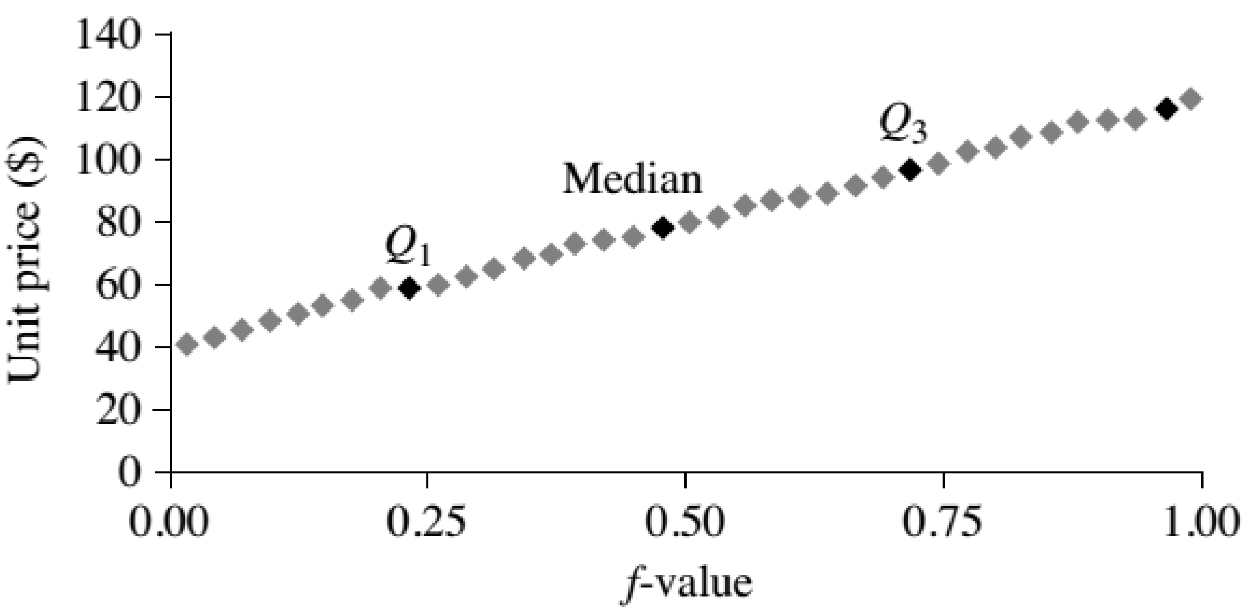
- A quantile plot presents the plot quantile information for a univariate data distribution
  - It allows access to both overall behavior and unusual occurrences.
- Let $x_1, x_2, \ldots, x_N$ be the data observations sorted in increasing order for some ordinal or numeric attribute X.
- Each value x_i is paired with $f_i = \frac{i-0.5}{N}$, indicating that approximately $f_i \times 100\%$ of data are $\leq x_i$.

Quantile plot: An example

A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350

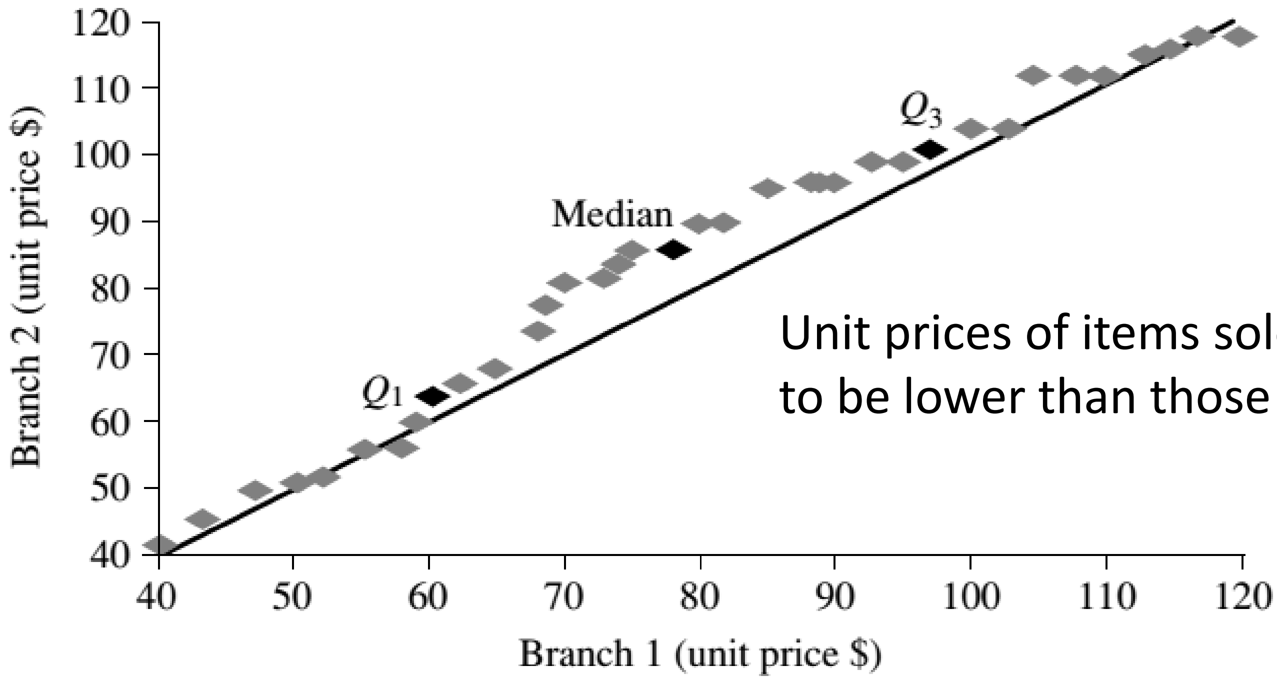
Quantile plot for the unit price data



Quantile-Quantile plot

- A quantile-quantile plot draws the quantiles of one univariate distribution against the corresponding quantiles of another.

Is there a shift in going from one distribution to another?



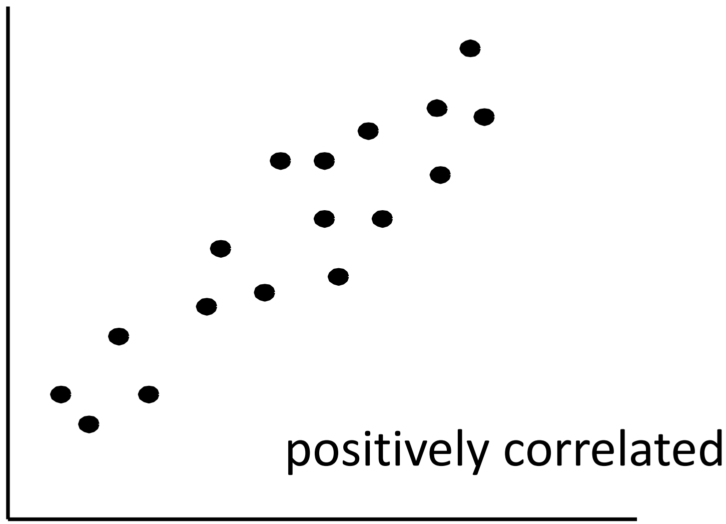
Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2

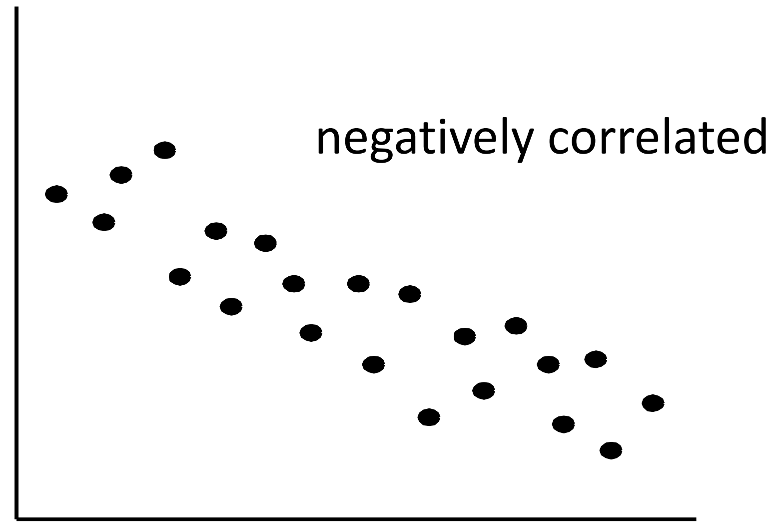
Scatter plot

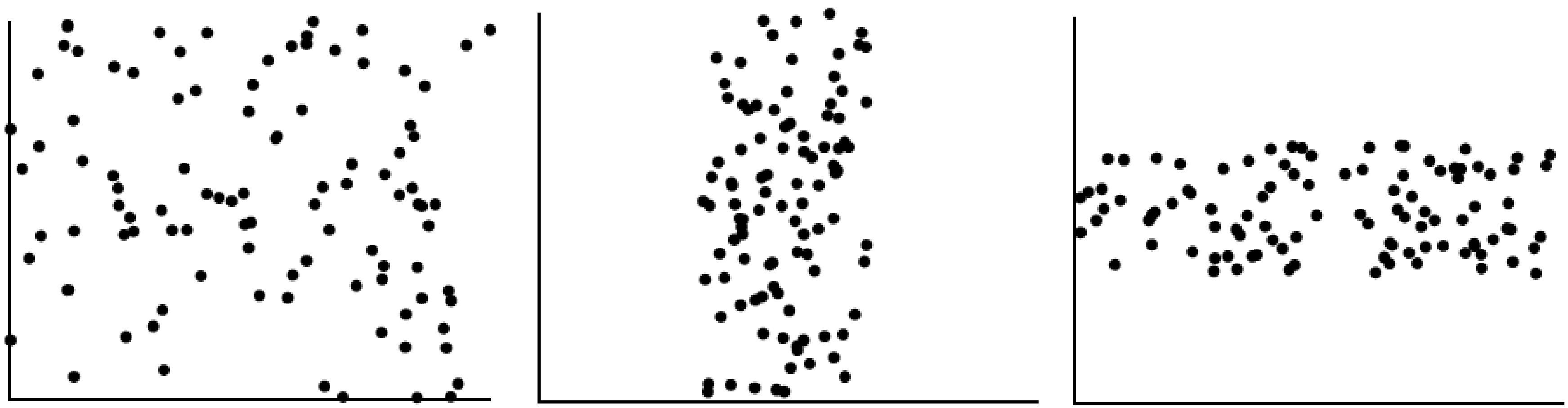
- A scatter plot looks at the bivariate data to see clusters of points or outliers
  - Each pair of values is treated as a pair of coordinates and plotted as points in the plane.

The correlation between Unit prices and Item sold

Scatter plot: Data correlation







uncorrelated data

Data proximity measures

Similarity and Dissimilarity

Similarity

- A numerical measure of how alike two data objects, i and j, are
- Values often falls in the range [0,1]: 0 – unalike $\rightarrow$ 1 – identical

Dissimilarity (distance)

- A numerical measure of how different two data objects are
- It works in an opposite direction to some similarity measure
- The lower bound is often 0, while the upper limit varies

Proximity

- This refers to either similarity or dissimilarity

Feature matrix vs. Dissimilarity matrix

- Feature matrices are essential to most machine learning task

Feature matrix

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

- n data points with p dimensions
- Object-by-attribute structure

Dissimilarity matrix

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

- A collection of distances for all pairs of n objects
- Object-by-object structure

- Many nearest-neighbor algorithms use dissimilarity matrices.

Measures for nominal attributes

- Let the number of states of a nominal attribute be M
- Method 1: Simple matching $\boldsymbol{d(i,j) = \frac{p-m}{p}}$
  - m: the number of attributes for which i and j are in the same state,
  - p: the total number of attributes describing the objects

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

$$\begin{bmatrix} 0 & & & \\ d(2,1) & 0 & & \\ d(3,1) & d(3,2) & 0 & \\ d(4,1) & d(4,2) & d(4,3) & 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

- Method 2: Create a binary attribute for each of the M states
- Measures of similarity $\boldsymbol{sim(i,j) = 1 - d(i,j) = \frac{m}{p}}$

Measures for binary attributes

- Contingency table

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

- Symmetric binary variable

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

- Asymmetric binary variable

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

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

Measures for binary attributes

name	*gender*	*fever*	*cough*	*test-1*	*test-2*	*test-3*	*test-4*
Jack	M	Y	N	P	N	N	N
Jim	M	Y	Y	N	N	N	N
Mary	F	Y	N	P	N	P	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

- Gender is symmetric binary, the remaining attributes are asymmetric
- Let the values Y and P be 1 and the value N be 0.
- Suppose that the distance between objects (patients) is computed based only on the asymmetric attributes
- $d(Jack, Jim) = \frac{1+1}{1+1+1} = 0.67,$ $d(Jack, Mary) = \frac{0+1}{2+0+1} = 0.33$

  $d(Jim, Mary) = \frac{1+2}{1+1+2} = 0.75$

Measures for numeric attributes

- Consider two data points of p-dimensional

$$i = (x_{i1}, x_{i2}, \dots, x_{ip}) \text{ and } j = (x_{j1}, x_{j2}, \dots, x_{ij})$$

- Minkowski distance (L_h norm)

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

  - where h is the order

Measures for numeric attributes

- $h = 1$: Manhattan (city block, L_1 norm) distance

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

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

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

- $h \to \infty$: "supremum" (L_{max} / L_∞ norm, Chebyshev) distance

$$d(i,j) = \lim_{h\to\infty}\left(\sum_{f=1}^{p}|x_{if} - x_{jf}|^h\right)^{1/h} = \max_f^p |x_{if} - x_{jf}|$$

Cosine similarity

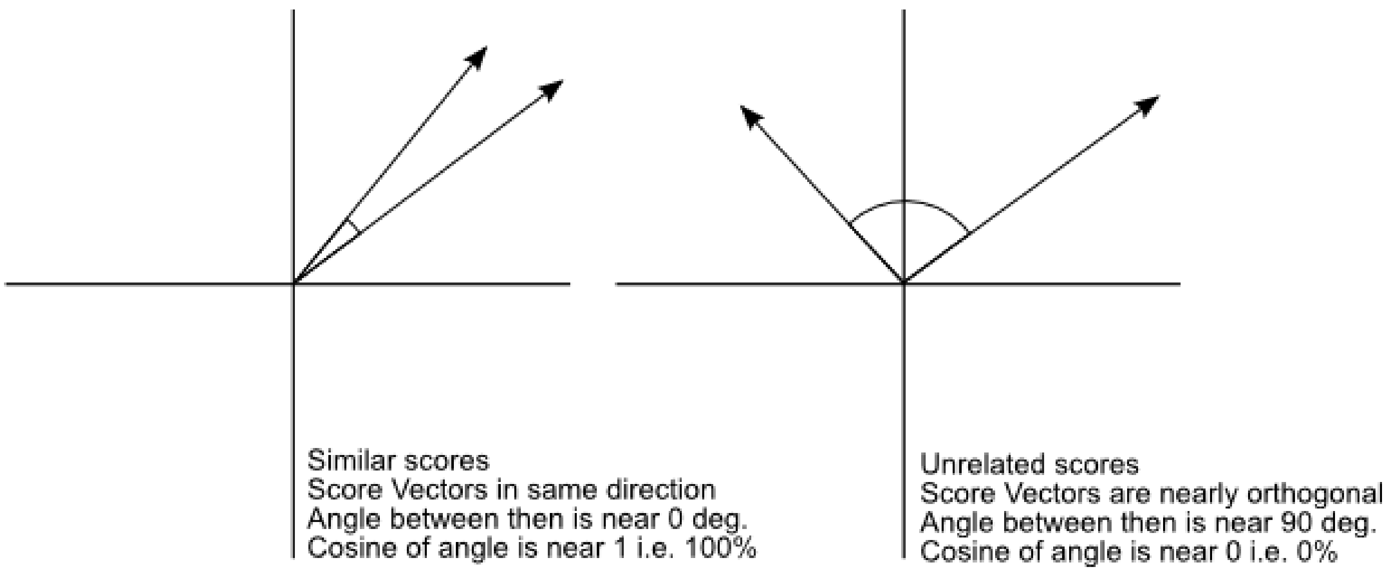
- A document can be represented by thousands of keywords in the document.

Document	team	coach	hockey	baseball	soccer	penalty	score	win	loss	season
Document1	5	0	3	0	2	0	0	2	0	0
Document2	3	0	2	0	1	1	0	1	0	1
Document3	0	7	0	[illegible]	[illegible]	[illegible]	0	3	0	0
Document4	0	1	0	0	1	2	2	0	3	0

$$sim(d_1, d_2) = 0.94$$

Cosine similarity

- Let d_1 and d_2 are two vectors (e.g., term-frequency vectors).
- Cosine similarity is non-metric: $sim(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\|\|d_3\|}$
  - where $\cdot$ is vector dot product, $\|d\|$ is the length of vector d
  - sim = 0 means no match, while sim = 1 means a complete match.



Measures for ordinal attributes

- The range of a numeric attribute can be mapped to an ordinal attribute f having M_f states.
  - E.g., temperate: cold (-30°C – 10°C), moderate (-10°C – 10°C), and warm (10°C – 30°C)
- Let M represent the number of possible ordered states, which define the ranking $1, \ldots, M_f$
- Replace each x_{if} by its corresponding rank, $r_{if} \in \{1, \ldots, M_f\}$
- Replace rank r_{if} of i^{th} object by $z_{if} = \dfrac{r_{if} - 1}{M_f - 1}$
- Continue with any measure for numeric attributes

Measures for ordinal attributes

Object Identifier	test-1 (nominal)	test-2 (ordinal)	test-3 (numeric)
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- test-2 = {fair, good, excellent}, i.e., $M_f = 3$
- The ranks of four objects are 3, 1, 2, and 3, respectively
- Map the rank 1 to 0.0, rank 2 to 0.5, and rank 3 to 1.0
- Dissimilarity matrix using Euclidean distance

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Measures for attributes of mixed types

- Suppose that the dataset has p attributes of mixed type.
- The distance between objects i and j is $\boldsymbol{d(i,j)} = \frac{\sum_{f=1}^{p} \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^{p} \delta_{ij}^{(f)}}$
  - $\delta_{ij}^{(f)} = 0$ if (1) x_{if} or x_{jf} is missing, or (2) $x_{if} = x_{jf} = 0$ and attribute f is asymmetric binary. Otherwise, $\delta_{ij}^{(f)} = 1$
  - If f is numeric: $d_{ij}^{(f)} = \frac{|x_{if} - x_{jf}|}{\max_h x_{hf} - \min_h x_{hf}}$, where h runs over all nonmissing objects for attribute f
  - If f is nominal or binary: $d_{ij}^{(f)} = 0$ if $x_{if} = x_{jf}$; otherwise, $d_{ij}^{(f)} = 1$
  - If f is ordinal: compute r_{if} and treat $z_{if} = \frac{r_{if} - 1}{M_f - 1}$ as numeric

Measures for attributes of mixed types

Dissimilarity matrix of test-1

$$\begin{bmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & 1 & 0 & \\ 0 & 1 & 1 & 0 \end{bmatrix}$$

Dissimilarity matrix of test-2

$$\begin{bmatrix} 0 & & & \\ 1.0 & 0 & & \\ 0.5 & 0.5 & 0 & \\ 0 & 1.0 & 0.5 & 0 \end{bmatrix}$$

Dissimilarity matrix of test-3

$$\begin{bmatrix} 0 & & & \\ 0.55 & 0 & & \\ 0.45 & 1.00 & 0 & \\ 0.40 & 0.14 & 0.86 & 0 \end{bmatrix}$$

Object Identifier	***test-1 (nominal)***	***test-2 (ordinal)***	***test-3 (numeric)***
1	code A	excellent	45
2	code B	fair	22
3	code C	good	64
4	code A	excellent	28

- $\delta_{ij}^{(f)} = 1$ for each attribute f
- $d(3,1) = \frac{1(1)+1(0.50)+1(0.45)}{3} = 0.65$
- The resulting dissimilarity matrix

$$\begin{bmatrix} 0 & & & \\ 0.85 & 0 & & \\ 0.65 & 0.83 & 0 & \\ 0.13 & 0.71 & 0.79 & 0 \end{bmatrix}$$

References

- Jiawei Han, Micheline Kamber, and Jian Pei. 2011. Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann Publishers Inc. Chapter 2.