Data Mining and Business Intelligence

ITA5007

PROF. E.P.EPHZIBAH

Topic

DIMENSION REDUCTION –DATA SUMMARIES

Dimension Reduction

The dimension of a dataset, which is the number of variables, must be reduced for the data mining algorithms to operate efficiently.

This process is part of the pilot/prototype phase of data mining and is done before deploying a model.

The dimensionality of a model is the number of predictors or input variables used by the model.

Curse of Dimensionality:

Key Idea "A function defined in high dimensional space is likely to be much more complex than a function defined in a lower-dimensional space, and those complications are harder to discern." —Milton Friedman (Famous Dude)

All about DATA

Data mining is the *process* of discovering interesting patterns and knowledge from *large* amounts of data. The data sources can include databases, data warehouses, the web, other information repositories, or data that are streamed into the system dynamically.

Data mining can also be applied to other forms of data like data streams, ordered/sequence data, graph or networked data, spatial data, text data, multimedia data, and the WWW.

Knowledge Discovery from Data, or KDD

Data mining is treated as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**.

The knowledge discovery process is an iterative sequence of the following steps:

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)

Knowledge Discovery from Data, or KDD

3. Data selection (where data relevant to the analysis task are retrieved from the database)

4. Data transformation (where data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)[4]

5. Data mining (an essential process where intelligent methods are applied to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on *interestingness measures*)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present mined knowledge to users)

Data set

Data sets are made up of data objects. A **data object** represents an entity—

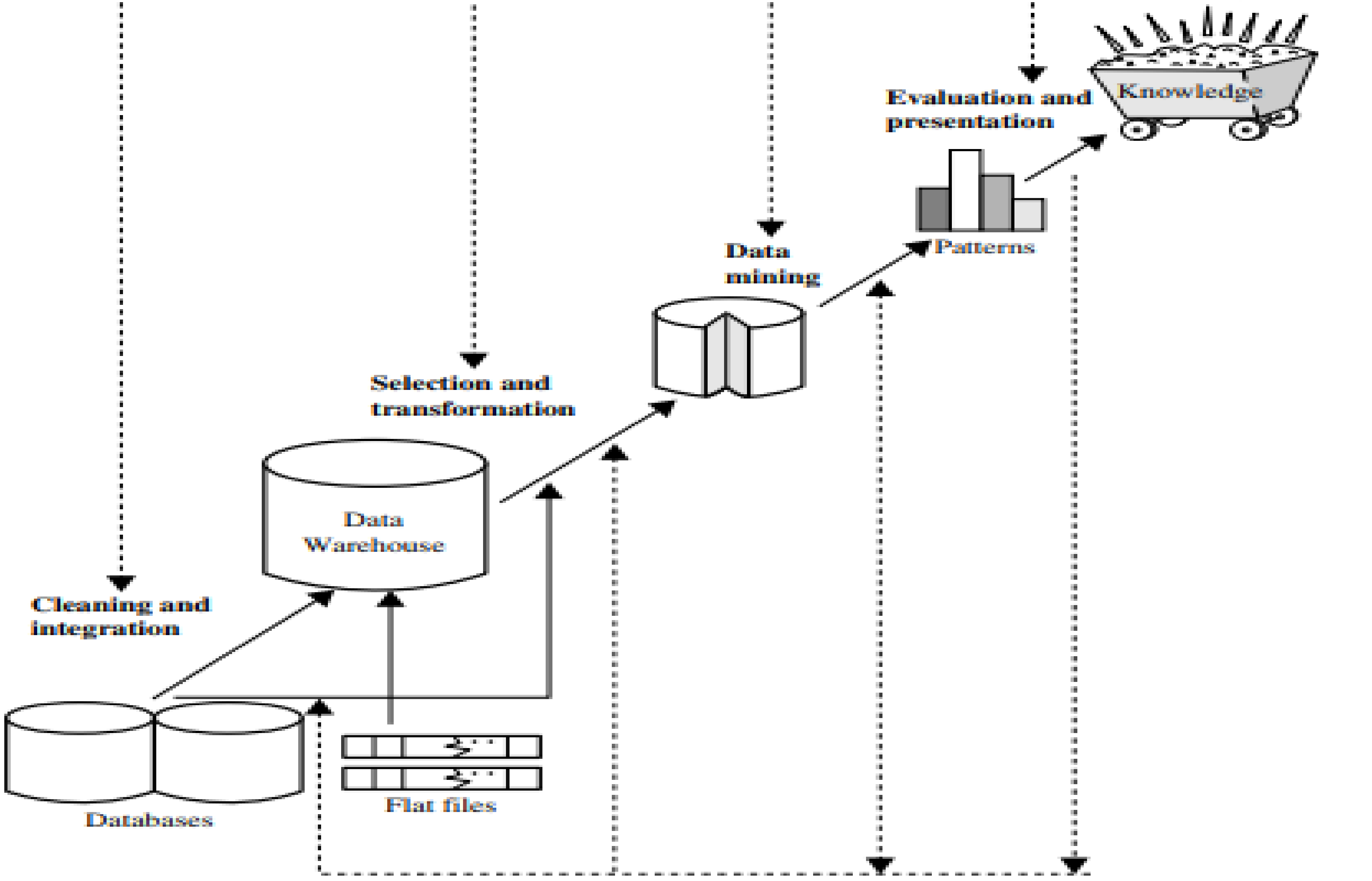
in a sales database, the objects may be customers, store items, and sales;

in a medical database, the objects may be patients;

in a university database, the objects may be students, professors, and courses.

Data objects are typically described by attributes. Data objects can also be referred to as *samples, examples, instances, data points*, or *objects*. If the data objects are stored in a database, they are *data tuples*.

KDD



Data mining as a step in the process of knowledge discovery.

What Is an Attribute?

An attribute is a data field, representing a characteristic or feature of a data object.

The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature.

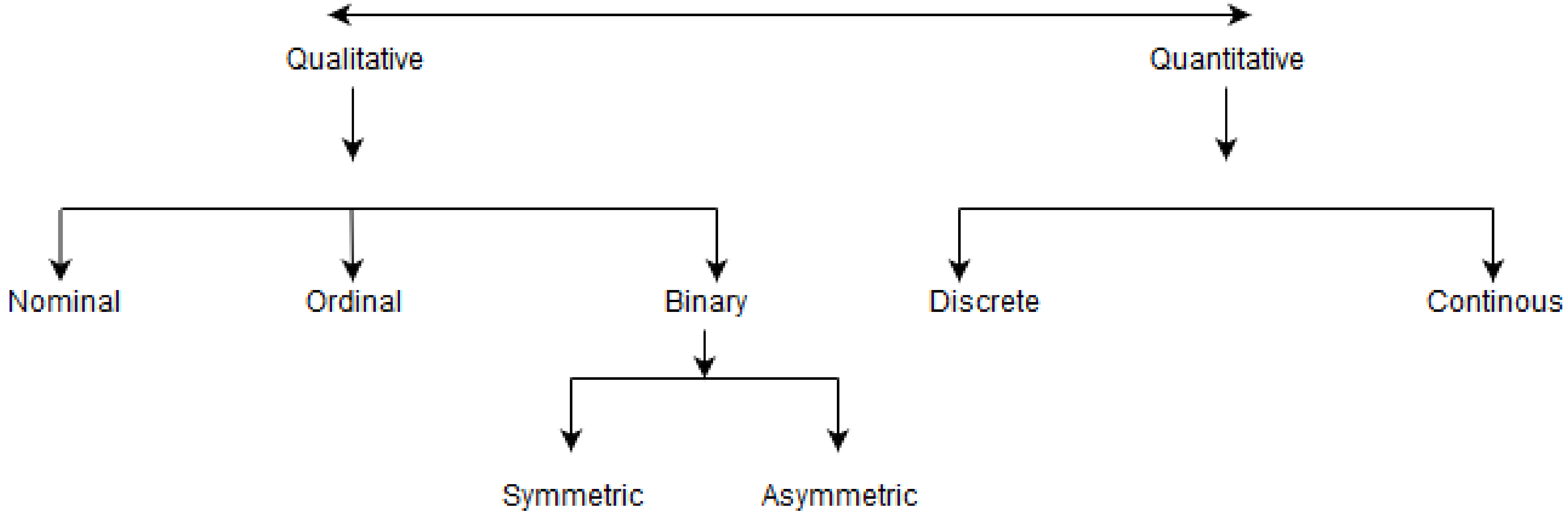
The type of an attribute is determined by the set of possible values—nominal, binary, ordinal, or numeric—the attribute can have.

Data

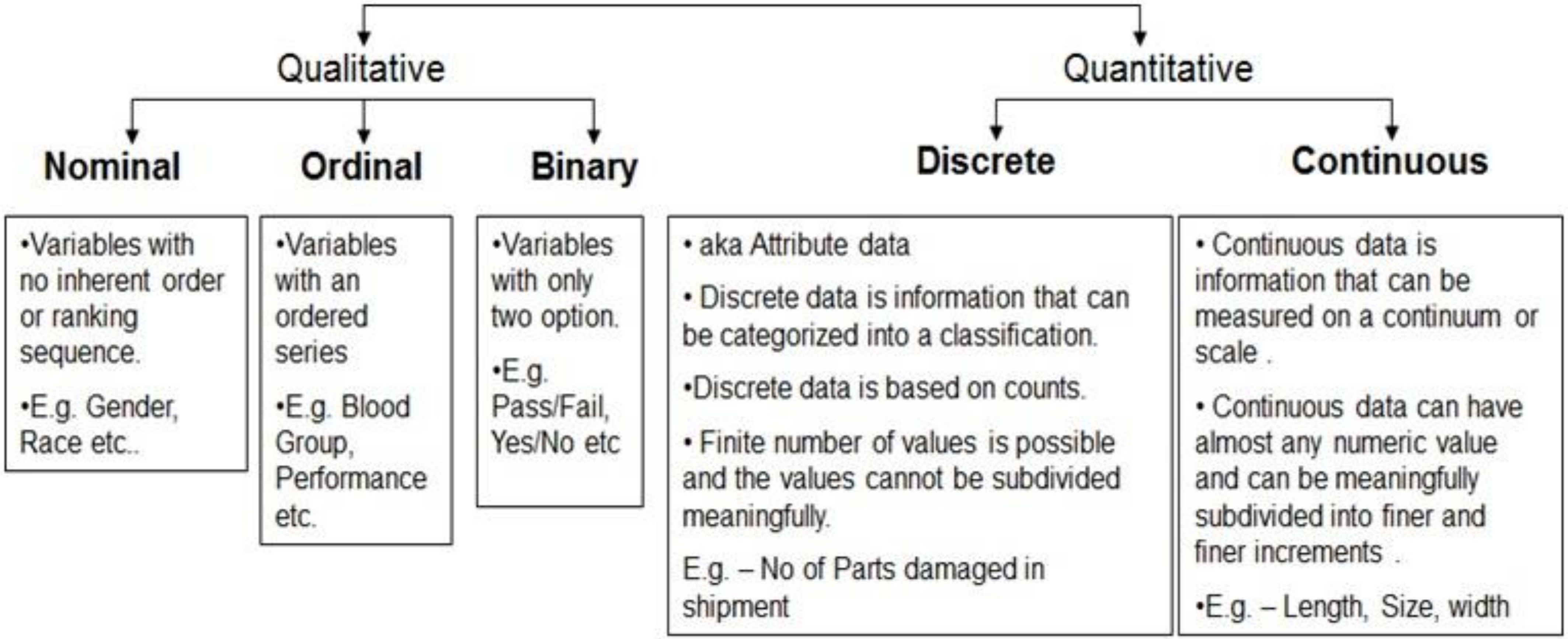
Types of Attributes

Attribute Type	Description	Examples
Nominal / Binary	The values are just different names that provide only enough information to distinguish one object from another. (=, ≠)	zip codes, employee ID numbers, eye color, gender
Ordinal	The values provide enough information to order objects. (<, >)	pain level, rating, grades, street numbers
Interval	The differences between values are meaningful, i.e., a unit of measurement exists (+, -)	calendar dates, temperature in Celsius or Fahrenheit
Ratio	Both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length

Data


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continous
Symmetric
Asymmetric

Data


Qualitative
Quantitative
Nominal
Ordinal
Binary
Discrete
Continuous
•Variables with no inherent order or ranking sequence.
•E.g. Gender, Race etc..
•Variables with an ordered series
•E.g. Blood Group, Performance etc.
•Variables with only two option.
•E.g. Pass/Fail, Yes/No etc
• aka Attribute data
• Discrete data is information that can be categorized into a classification.
•Discrete data is based on counts.
• Finite number of values is possible and the values cannot be subdivided meaningfully.
E.g. – No of Parts damaged in shipment
• Continuous data is information that can be measured on a continuum or scale .
• Continuous data can have almost any numeric value and can be meaningfully subdivided into finer and finer increments .
•E.g. – Length, Size, width

Nominal Attributes

Nominal means “relating to names.” The values of a **nominal attribute** are symbols or *names of things*. Each value represents some kind of category, code, or state, and so nominal attributes are also referred to as **categorical**.

The values do not have any meaningful order. In computer science, the values are also known as *enumerations*.

The attribute *marital status* can take on the values *single, married, divorced*, and *widowed*.

Another example of a nominal attribute is *occupation*, with the values *teacher, dentist, programmer, farmer*, and so on.

Binary Attributes

A binary attribute is a nominal attribute with only two categories or states: 0 or 1, where 0 typically means that the attribute is absent, and 1 means that it is present. Binary attributes are referred to as Boolean if the two states correspond to true and false.

Given the attribute player describing a person object, 1 indicates that the person is a player , while 0 indicates that the person is not a player.

Ordinal Attributes

An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them.

Suppose that drink size corresponds to the size of drinks available at a fast-food restaurant. This nominal attribute has three possible values: small, medium, and large.

Other examples of ordinal attributes include grade (e.g., A++, A+, A, B, and so on)

Customer satisfaction had the following ordinal categories:

0: very dissatisfied, 1: somewhat dissatisfied, 2: neutral, 3: satisfied, and *4: very satisfied.*

Numeric Attributes

A numeric attribute is quantitative; that is, it is a measurable quantity, represented in integer or real values. Numeric attributes can be interval-scaled or ratio-scaled.

- Interval-scaled attributes are measured on a scale of equal-size units. The values of interval-scaled attributes have order and can be positive, 0, or negative.
- A ratio-scaled attribute is a numeric attribute with an inherent zero-point. That is, if a measurement is ratio-scaled, we can speak of a value as being a multiple (or ratio) of another value.

Interval-scaled attributes

Temperature attribute is interval scaled.

Suppose that we have the outdoor temperature value for a number of different days, where each day is an object. By ordering the values, we obtain a ranking of the objects with respect to temperature.

Examples of ratio-scaled attributes include count attributes such as years of experience (e.g., the objects are employees) and number of words (e.g., the objects are documents).

Discrete and continuous attributes

Discrete attribute has a finite or countable infinite set of values, which may or may not be represented as integers. The attributes *hair colour*, *smoker*, *medical test*, and *drink size* each have a finite number of values, and so are discrete.

If an attribute is not discrete, it is **continuous**. The terms *numeric attribute* and *continuous attribute* are often used interchangeably in the literature.

Exercise 1:

In the house price predition dataset , identify the type of data for each and every attribute. Prepare a document for one page.

House Price Prediction dataset

date	price	bedrooms	bathroom	sqft_living	sqft_lot	floors	waterfron	view	condition	sqft_abov	sqft_base	yr_built	yr_renova	street	city	statezip	country
02-05-2014 00:00	313000	3	1.5	1340	7912	1.5	0	0	3	1340	0	1955	2005	18810 Den	Shoreline	WA 98133	USA
02-05-2014 00:00	2384000	5	2.5	3650	9050	2	0	4	5	3370	280	1921	0	709 W Blai	Seattle	WA 98119	USA
02-05-2014 00:00	342000	3	2	1930	11947	1	0	0	4	1930	0	1966	0	26206-262	Kent	WA 98042	USA
02-05-2014 00:00	420000	3	2.25	2000	8030	1	0	0	4	1000	1000	1963	0	857 170th	Bellevue	WA 98008	USA
02-05-2014 00:00	550000	4	2.5	1940	10500	1	0	0	4	1140	800	1976	1992	9105 170th	Redmond	WA 98052	USA
02-05-2014 00:00	490000	2	1	880	6380	1	0	0	3	880	0	1938	1994	522 NE 88t	Seattle	WA 98115	USA
02-05-2014 00:00	335000	2	2	1350	2560	1	0	0	3	1350	0	1976	0	2616 174th	Redmond	WA 98052	USA
02-05-2014 00:00	482000	4	2.5	2710	35868	2	0	0	3	2710	0	1989	0	23762 SE 2	Maple Val	WA 98038	USA
02-05-2014 00:00	452500	3	2.5	2430	88426	1	0	0	4	1570	860	1985	0	46611-466	North Ben	WA 98045	USA
02-05-2014 00:00	640000	4	2	1520	6200	1.5	0	0	3	1520	0	1945	2010	6811 55th	Seattle	WA 98115	USA
02-05-2014 00:00	463000	3	1.75	1710	7320	1	0	0	3	1710	0	1948	1994	Burke-Gilr	Lake Fores	WA 98155	USA
02-05-2014 00:00	1400000	4	2.5	2920	4000	1.5	0	0	5	1910	1010	1909	1988	3838-4098	Seattle	WA 98105	USA
02-05-2014 00:00	588500	3	1.75	2330	14892	1	0	0	3	1970	360	1980	0	1833 220th	Sammami	WA 98074	USA
02-05-2014 00:00	365000	3	1	1090	6435	1	0	0	4	1090	0	1955	2009	2504 SW P	Seattle	WA 98106	USA
02-05-2014 00:00	1200000	5	2.75	2910	9480	1.5	0	0	3	2910	0	1939	1969	3534 46th	Seattle	WA 98105	USA
02-05-2014 00:00	242500	3	1.5	1200	9720	1	0	0	4	1200	0	1965	0	14034 SE 2	Kent	WA 98042	USA
02-05-2014 00:00	419000	3	1.5	1570	6700	1	0	0	4	1570	0	1956	0	15424 SE 9	Bellevue	WA 98007	USA
02-05-2014 00:00	367500	4	3	3110	7231	2	0	0	3	3110	0	1997	0	11224 SE 3	Auburn	WA 98092	USA
02-05-2014 00:00	257950	3	1.75	1370	5858	1	0	0	3	1370	0	1987	2000	1605 S 245	Des Moine	WA 98198	USA
02-05-2014 00:00	275000	3	1.5	1180	10277	1	0	0	3	1180	0	1983	2009	12425 415t	North Ben	WA 98045	USA
02-05-2014 00:00	750000	3	1.75	2240	10578	2	0	0	5	1550	690	1923	0	3225 NE 92	Seattle	WA 98115	USA
02-05-2014 00:00	435000	4	1	1450	8800	1	0	0	4	1450	0	1954	1979	3922 154th	Bellevue	WA 98006	USA

Data Preprocessing Techniques

Data cleaning can be applied to remove noise and correct inconsistencies in data.

Data integration merges data from multiple sources into a coherent data store such as a data warehouse.

Data reduction can reduce data size by, for instance, aggregating, eliminating redundant features, or clustering.

Data transformations (e.g., normalization) may be applied, where data are scaled to fall within a smaller range like 0.0 to 1.0.

Impacts of data preprocessing

There are many factors comprising

data quality,

accuracy,

completeness,

consistency,

timeliness,

believability, and

interpretability.

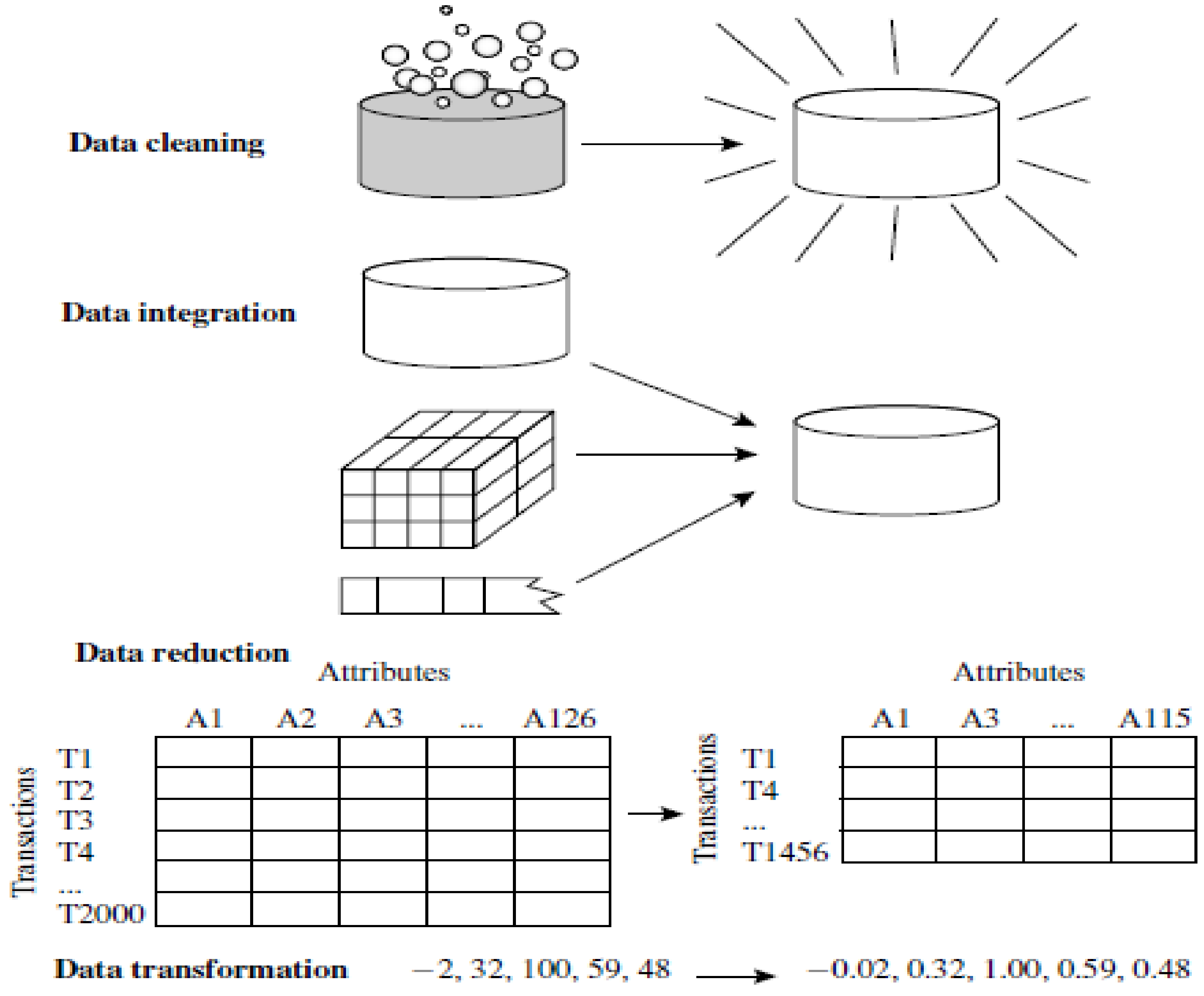


Figure 3.1 Forms of data preprocessing.

DATA CLEANING

Methods to handle Missing Values:

- Ignore the tuple
- Fill in the missing value manually
- Use a global constant to fill in the missing value
- Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value
- Use the attribute mean or median for all samples belonging to the same class as the given tuple
- Use the most probable value to fill in the missing value (using prediction methods)

NOISY DATA

"What is noise?" Noise is a random error or variance in a measured variable.

Given a numeric attribute such as, say, price, how can we "smooth" out the data to remove the noise?

Let's look at the following **data smoothing techniques**.

- **Binning:** Binning methods smooth a sorted data value by consulting its "neighbourhood," that is, the values around it. The sorted values are distributed into a number of "buckets," or bins. Because binning methods consult the neighbourhood of values, they perform local smoothing.

Sorted data for *price* (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Partition into (equal-frequency) bins:

Bin 1: 4, 8, 15
Bin 2: 21, 21, 24
Bin 3: 25, 28, 34

Smoothing by bin means:

Bin 1: 9, 9, 9
Bin 2: 22, 22, 22
Bin 3: 29, 29, 29

Smoothing by bin boundaries:

Bin 1: 4, 4, 15
Bin 2: 21, 21, 24
Bin 3: 25, 25, 34

Data smoothing techniques:

- Regression:
  - *Linear regression* involves finding the “best” line to fit two attributes (or variables) so that one attribute can be used to predict the other. *Multiple linear regression* is an extension of linear regression, where more than two attributes are involved and the data are fit to a multidimensional surface.
- Outlier analysis:
  - Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.” Intuitively, values that fall outside of the set of clusters may be considered outliers

Data clusters

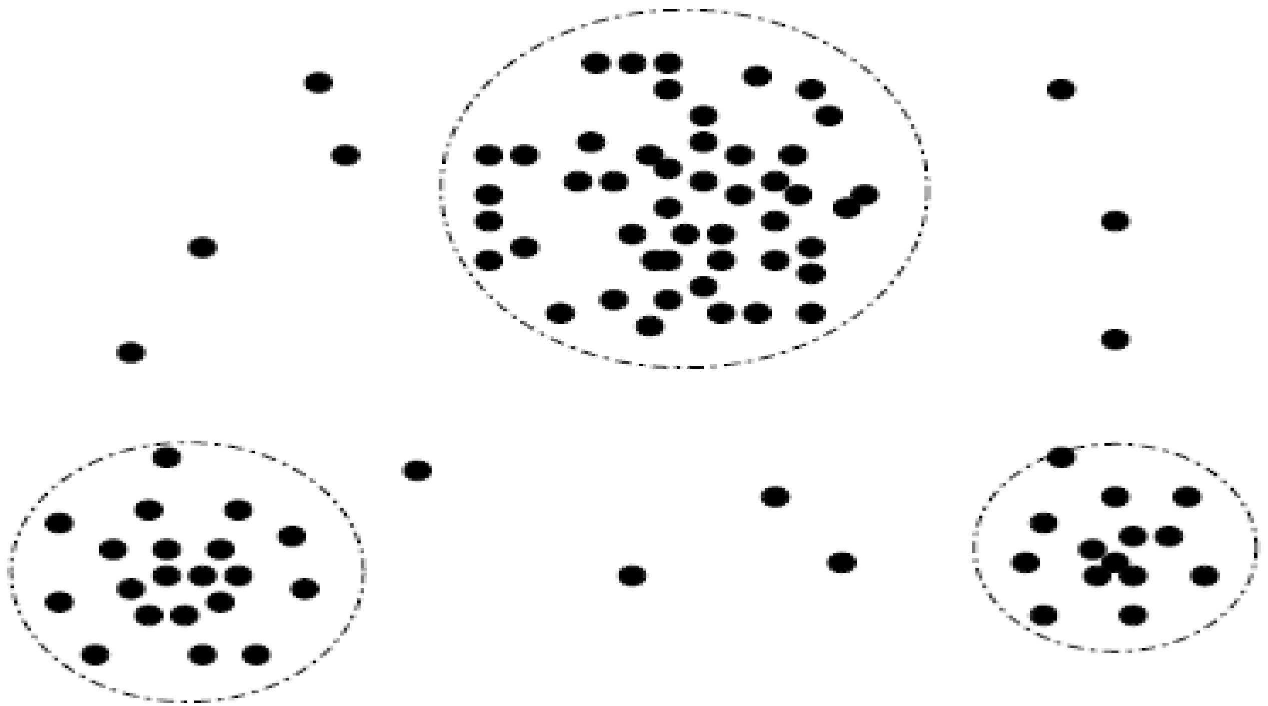


Figure 3.3 A 2-D customer data plot with respect to customer locations in a city, showing three data clusters. Outliers may be detected as values that fall outside of the cluster sets.

Data reduction startegy -Histogram

Histograms use binning to approximate data distributions and are a popular form of data reduction.

A histogram for an attribute, A, partitions the data distribution of A into disjoint subsets, referred to as buckets or bins.

If each bucket represents only a single attribute–value/frequency pair, the buckets are called singleton buckets. Often, buckets instead represent continuous ranges for the given attribute.

Example - Histograms

The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

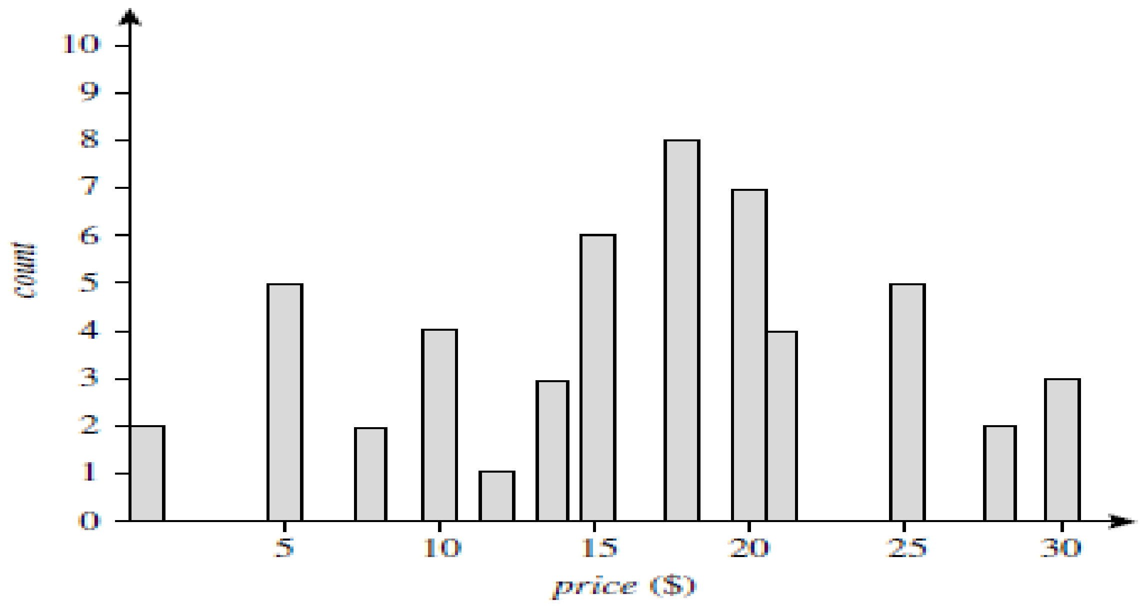


Figure 3.7 A histogram for *price* using singleton buckets—each bucket represents one price–value/frequency pair.

Data transformation strategies

In data transformation, the data are transformed or consolidated into forms appropriate for mining.

1.**Smoothing**, which works to remove noise from the data. Techniques include binning, regression, and clustering.

2. **Attribute construction** (or feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

3. **Aggregation**, where summary or aggregation operations are applied to the data. For example, the daily sales data may be aggregated so as to compute monthly and annual total amounts.

4. **Normalization**, where the attribute data are scaled so as to fall within a smaller range, such as 1.0 to 1.0, or 0.0 to 1.0.

5. **Discretization**, where the raw values of a numeric attribute (e.g., age) are replaced by interval labels (e.g., 0–10, 11–20, etc.) or conceptual labels (e.g., youth, adult, senior).

6. **Concept hierarchy** generation for nominal data, where attributes such as street can be generalized to higher-level concepts, like city or country.

Data Transformation by Normalization

Normalizing the data attempts to give all attributes an equal weight.

Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbour classification and clustering.

Min-max normalization

Min-max normalization performs a linear transformation on the original data. Suppose that min_A and max_A are the minimum and maximum values of an attribute, A. Min-max normalization maps a value, v_i, of A to v_i' in the range $[new\_min_A, new\_max_A]$ by computing

$$v_i' = \frac{v_i - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A. \tag{3.8}$$

Min-max normalization preserves the relationships among the original data values. It will encounter an "out-of-bounds" error if a future input case for normalization falls outside of the original data range for A.

Min-max normalization. Suppose that the minimum and maximum values for the attribute *income* are \$12,000 and \$98,000, respectively. We would like to map *income* to the range [0.0, 1.0]. By min-max normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600-12,000}{98,000-12,000}(1.0-0)+0 = 0.716$. ■

Z-SCORE NORMALIZATION

In **z-score normalization** (or *zero-mean normalization*), the values for an attribute, A, are normalized based on the mean (i.e., average) and standard deviation of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i - \bar{A}}{\sigma_A}, \tag{3.9}$$

where $\bar{A}$ and σ_A are the mean and standard deviation, respectively, of attribute A. The

z-score normalization. Suppose that the mean and standard deviation of the values for the attribute *income* are \$54,000 and \$16,000, respectively. With z-score normalization, a value of \$73,600 for *income* is transformed to $\frac{73,600 - 54,000}{16,000} = 1.225$. ■

Decimal Scaling

Normalization by decimal scaling normalizes by moving the decimal point of values of attribute A. The number of decimal points moved depends on the maximum absolute value of A. A value, v_i, of A is normalized to v_i' by computing

$$v_i' = \frac{v_i}{10^j}, \tag{3.12}$$

where j is the smallest integer such that $max(|v_i'|) < 1$.

Decimal scaling. Suppose that the recorded values of A range from −986 to 917. The maximum absolute value of A is 986. To normalize by decimal scaling, we therefore divide each value by 1000 (i.e., $j = 3$) so that −986 normalizes to −0.986 and 917 normalizes to 0.917. ■

Data summaries

Today's real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size (often several gigabytes or more) and their likely origin from multiple, heterogenous sources.

Low-quality data will lead to low-quality mining

results.

"How can the data be preprocessed in order to help improve the quality of the data and, consequently, of the mining results?

How can the data be preprocessed so as to improve the efficiency and ease of the mining process?"

Mean

Mean. Suppose we have the following values for salary (in lakhs), shown in increasing order:

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}.$$

Mean

$$\bar{x} = \frac{30+36+47+50+52+52+56+60+63+70+70+110}{12}$$

$$= \frac{696}{12} = 58.$$

Thus, the mean salary is 58 lakhs

Median

30, 36, 47, 50, 52, 56, 60, 63, 70, 70, 110 (odd number of entries)

Median= element in 6th position 56

Value	30	36	47	50	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110 (Even number of entries)

Value	30	36	47	50	52	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11	12

Median= elements in 5th and 6th position 52 and 56

$$\frac{52+56}{2} = \frac{108}{2} = 54.$$

Mode

Data: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110.

Here mode =52, 70 (occurs 2 times) - Bimodal

For unimodal numeric data that are moderately skewed (asymmetrical), we have the following empirical relation:

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

Midrange

The midrange can also be used to assess the central tendency of a numeric data set.

It is the average of the largest and smallest values in the set.

30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

Midrange =30+110 /2 =70

Range

The range of the set is the difference between the largest (max()) and smallest (min()) values.

Data: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110

Range =110-30 =80

Quantile

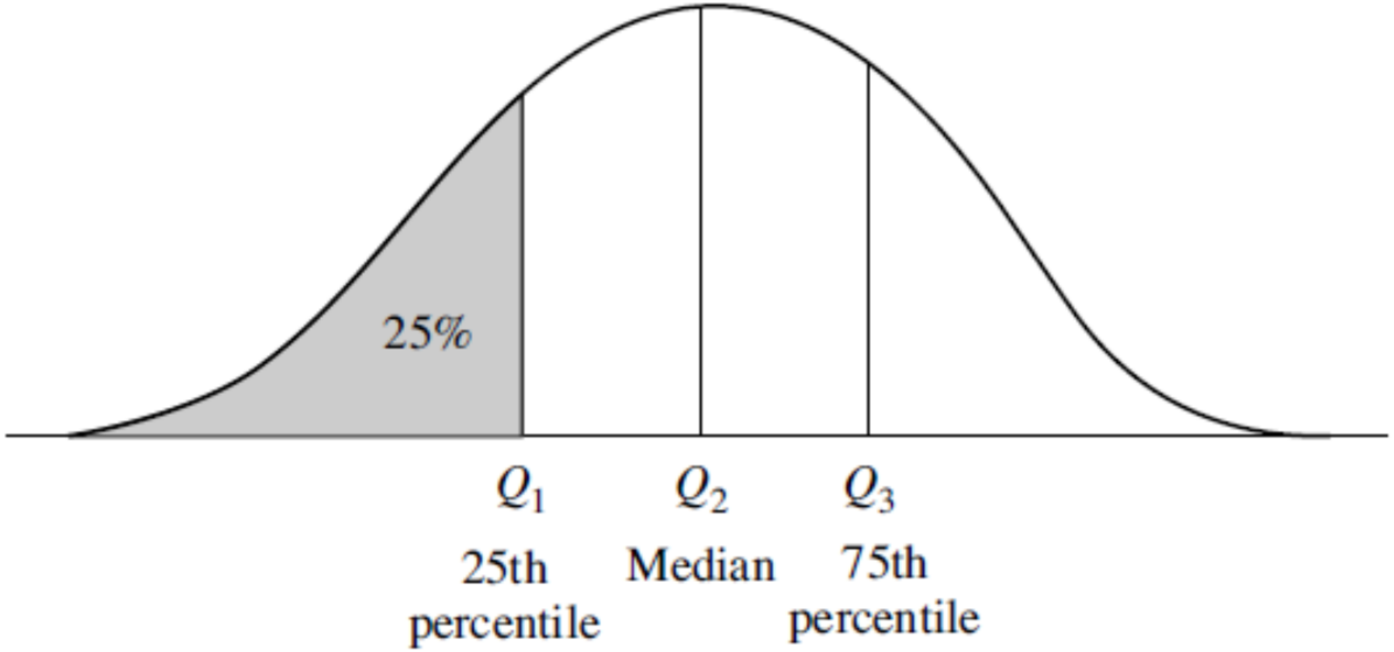
Quantile means one of the classes of values of a variable which divides the members of a batch or sample into equal-sized subgroups of adjacent values or a probability distribution into distributions of equal probability

The data points that split the data distribution into equal-sized consecutive sets are called *quantiles*.

Quantiles are points taken at regular intervals of data distribution, dividing it into essentially equal size consecutive sets.

Quartile

Quartile means any of the three points that divide an ordered distribution into four parts, each containing a quarter of the population.



Types of quantiles

The 2-quantile is the data point dividing the lower and upper halves of the data distribution. It corresponds to the median.

The 4-quantiles are the three data points that split the data distribution into four equal parts; each part represents one-fourth of the data distribution. They are more commonly referred to as quartiles.

The 100-quantiles are more commonly referred to as percentiles; they divide the data distribution into 100 equal-sized consecutive sets.

The median, quartiles, and percentiles are the most widely used forms of quantiles.

Inter Quartile Range (IQR)

The distance between the first and third quartiles is a simple measure of spread that gives the range covered by the middle half of the data. This distance is called the interquartile range (IQR) and is defined as

IQR = Q3 -Q1

Value	30	36	47	50	52	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11	12

There are 12 observations, already sorted in increasing order.

The quartiles for this data are the third, sixth, and ninth values, respectively, in the sorted list. Therefore, Q1 is 47 and Q3 is 63.

Thus, the interquartile range is IQR = 63-47 =16

Five number summary

The five-number summary of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of

Minimum, Q1, Median, Q3, Maximum

Box plots

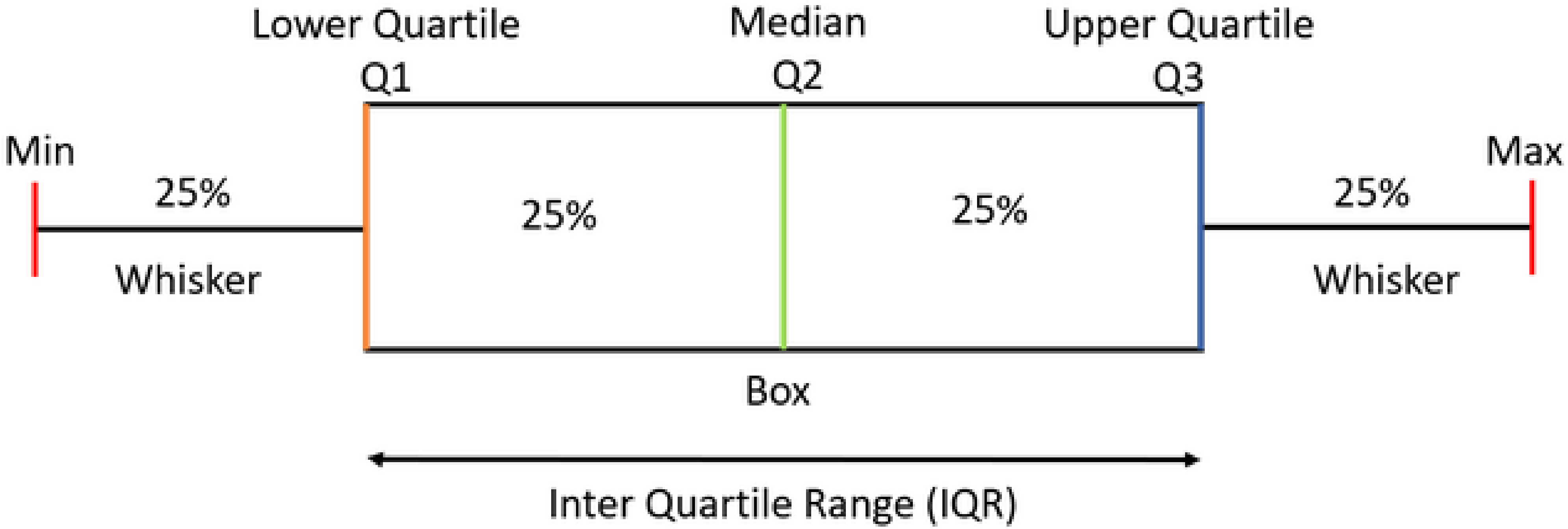
Boxplots are a popular way of visualizing a distribution. A boxplot incorporates the five-number summary as follows:

The ends of the box are at the quartiles so that the box length is the interquartile range.

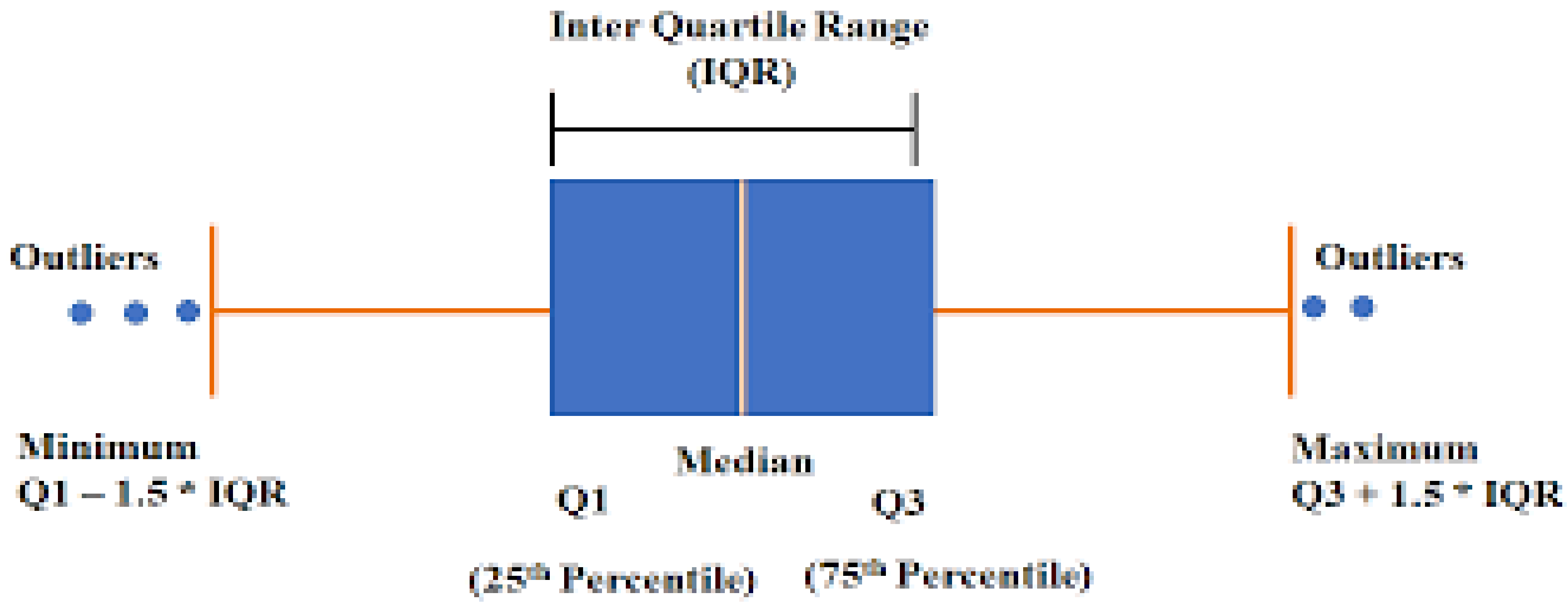
The median is marked by a line within the box.

Two lines (called whiskers) outside the box extend to the smallest (Minimum) and largest (Maximum) observations.

Box plot


Lower Quartile
Q1
Median
Q2
Upper Quartile
Q3
Min
Max
25%
25%
25%
25%
Whisker
Whisker
Box
Inter Quartile Range (IQR)

Outlier Identification



Value	30	36	47	50	52	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11	12

Five number summary

Min =30

Q1=47

Median=52

Q3=63

IQR= 16

Q1-1.5 X IQR=47-1.5 X 16= 47-24 =23 any value below 23 is an outlier

Q3+1.5 X IQR=63+1.5 X 16= 63+24 =87 any value above 87 is an outlier

Draw the box plot

Value	30	36	47	50	52	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11	12

Variance and Standard Deviation

Variance and standard deviation are measures of data dispersion.

They indicate how spread out a data distribution is.

A low standard deviation means that the data observations tend to be very close to the mean, while a high standard deviation indicates that the data are spread out over a large range of values.

Variance

The **variance** of N observations, $x_1, x_2, \ldots, x_N$, for a numeric attribute X is

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

where x is the mean value of the observations.

The standard deviation, σ, of the observations is the square root of the variance.

Value	30	36	47	50	52	52	56	60	63	70	70	110
Position	1	2	3	4	5	6	7	8	9	10	11	12

Variance
$$\sigma^2 = \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2$$
$$\approx 379.17$$

Standard Deviation
$$\sigma \approx \sqrt{379.17} \approx 19.47.$$

Quantile Plot

A quantile plot is a simple and effective way to have a first look at a univariate data distribution.

First, it displays all of the data for the given attribute

Second, it plots quantile information

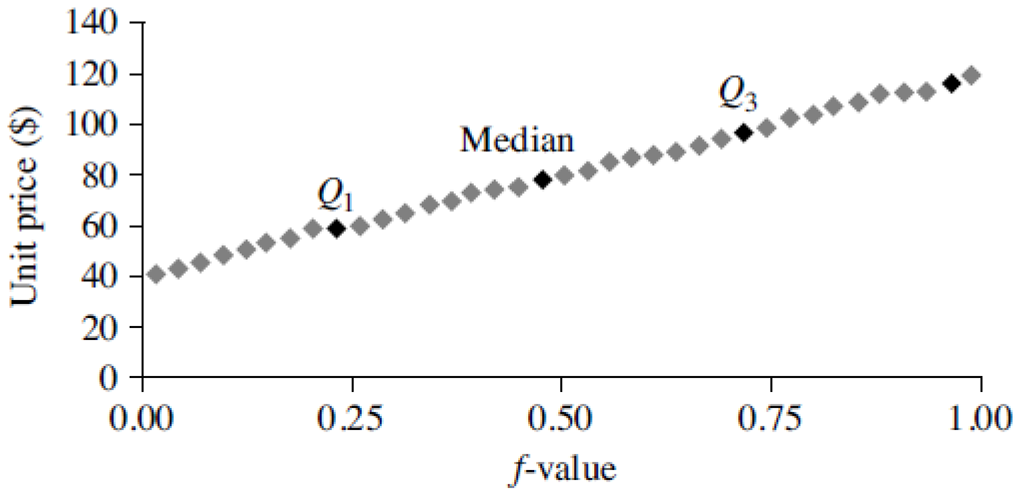
On a quantile plot, *xi* is graphed against *fi* .

$$f_i = \frac{i - 0.5}{N}$$

Quantile Plot

Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

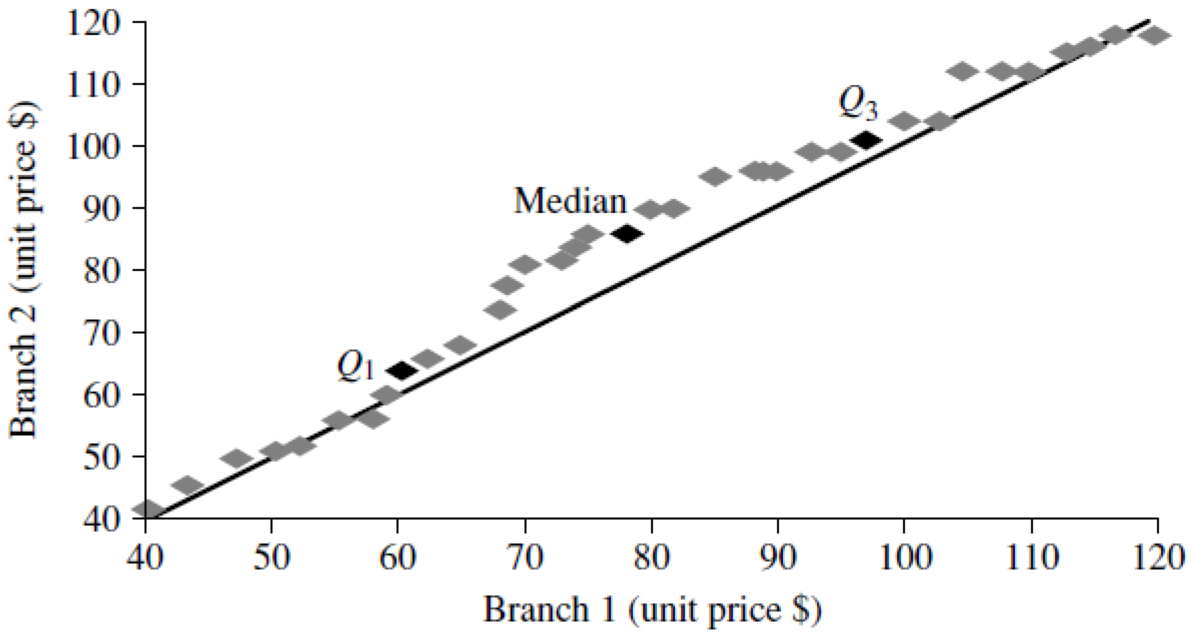
Unit price ($)	*Count of items sold*
40	275
43	300
47	250
—	—
74	360
75	515
78	540
—	—
115	320
117	270
120	350



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

Quantile–Quantile Plot or q-q plot

Aquantile–quantile plot, or q-q plot, graphs the quantiles of one univariate distribution against the corresponding quantiles of another.



Histograms

Histogram is a chart of poles.

Plotting histograms is a graphical method for summarizing the distribution of a given attribute, *X*.

If *X* is nominal, such as *automobile model* or *item type*, then a pole or vertical bar is drawn for each known value of *X*. The height of the bar indicates the frequency (i.e., count) of that *X* value. The resulting graph is more commonly known as a **bar chart**.

Histograms

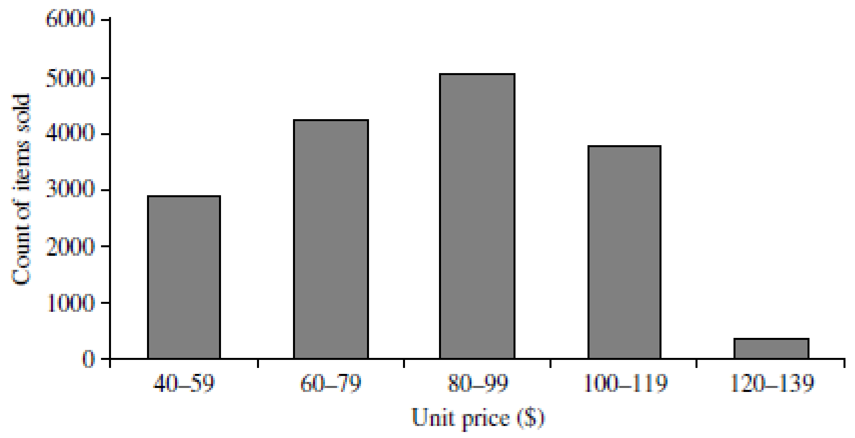
If X is numeric, the term histogram is preferred.

The range of values for X is partitioned into disjoint consecutive subranges.

The subranges, referred to as buckets or bins, are disjoint subsets of the data distribution for X. The range of a bucket is known as the width.

Table 2.1 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



A histogram for the Table 2.1 data set.

Scatter Plots

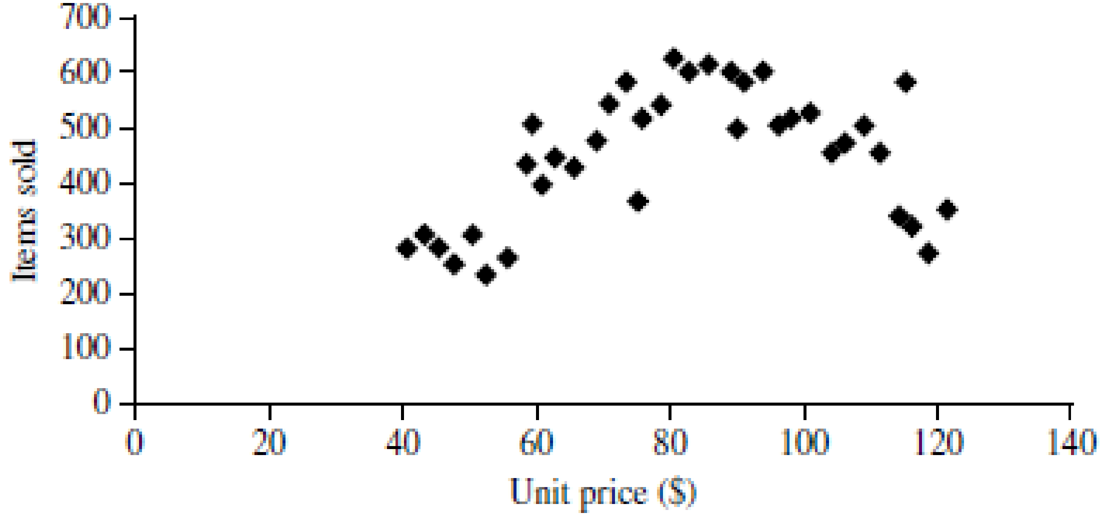
A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, pattern, or trend between two numeric attributes.

The scatter plot is a useful method for providing a first look at bivariate data to see clusters of points and outliers, or to explore the possibility of correlation relationships.

Scatter plot

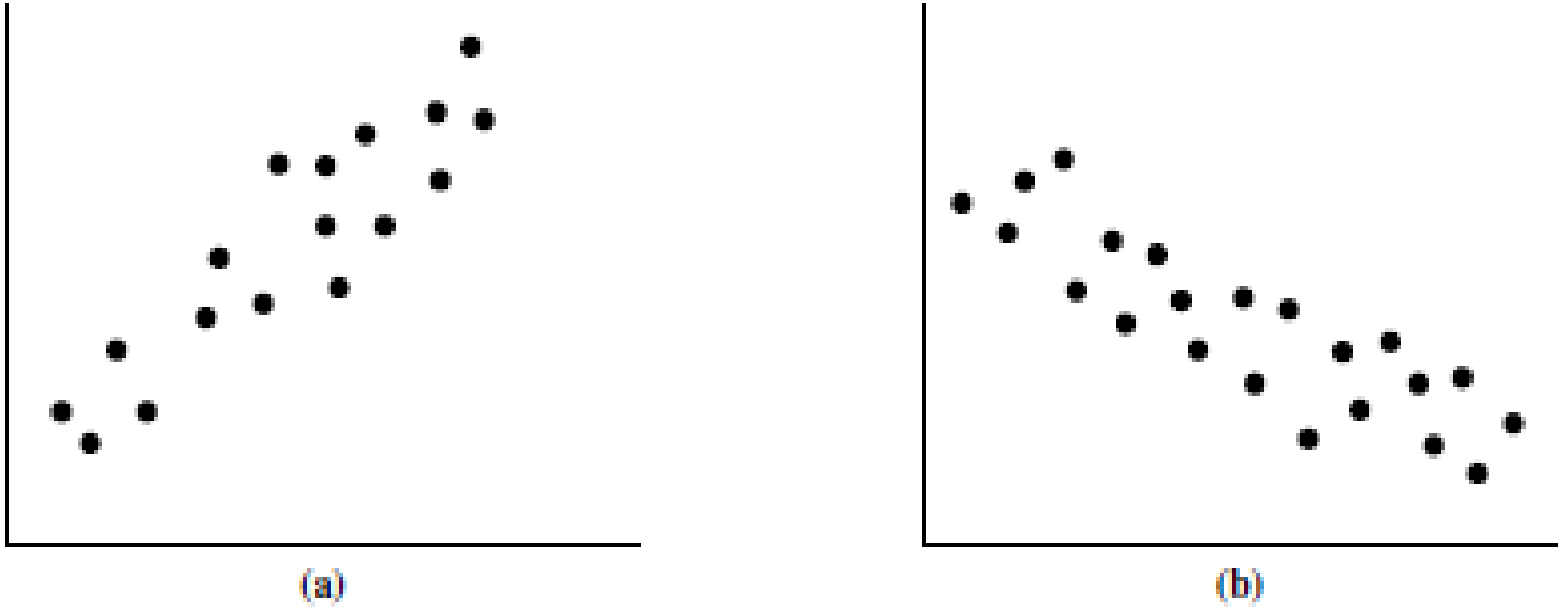
Table 2.1 A Set of Unit Price Data for Items Sold at a Branch of *AllElectronics*

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



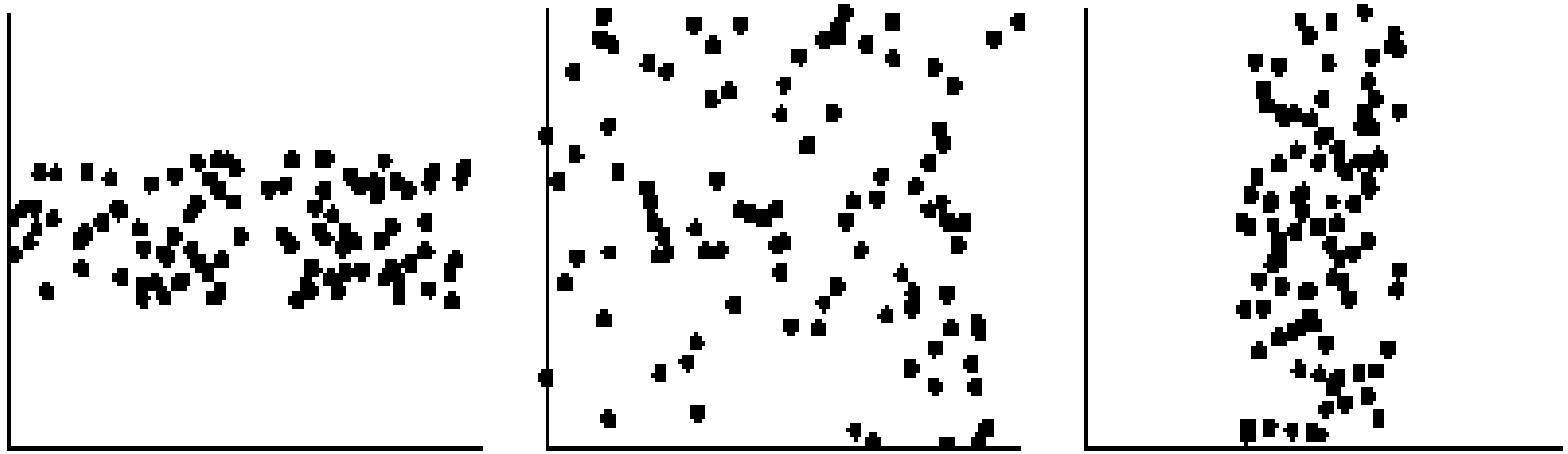
A scatter plot for the Table 2.1 data set.

Scatter plot and correlation



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

Types of correlation



Three cases where there is no observed correlation between the two plotted attributes in each of the data sets.