



# **FUNDAMENTALS OF MACHINE LEARNING IN DATA SCIENCE**

**CSIS 3290**

**DATA PREPROCESSING**

**FATEMEH AHMADI**



# Data Quality: Why Preprocess the Data?

Data have quality if they satisfy the requirements of the intended use. There are many factors comprising data quality (DQ), including:

- Accuracy,
- Completeness,
- Consistency,
- Timeliness,
- Believability,
- Interpretability



# Accuracy, completeness, and consistency

Inaccurate, incomplete, and inconsistent data are commonplace properties of large real-world databases and data warehouses. There are many possible reasons for inaccurate data (i.e., having incorrect attribute values).

- The data collection instruments used may be faulty.
- There may have been human or computer errors occurring at data entry.
- Users may purposely submit incorrect data values for mandatory fields when they do not wish to submit personal information (e.g., by choosing the default value “January 1” displayed for birthday). This is known as **disguised missing data**.
- Errors in data transmission can also occur.
- There may be technology limitations such as limited buffer size for coordinating synchronized data transfer and consumption.
- Incorrect data may also result from inconsistencies in naming conventions or data codes, or inconsistent formats for input fields (e.g., date).
- Duplicate tuples also require data cleaning.



# Believability and interpretability

- Two other factors affecting data quality are believability and interpretability.
- **Believability** reflects how much the data are trusted by users, while **interpretability** reflects how easy the data are understood.
- Suppose that a database, at one point, had several errors, all of which have since been corrected. The past errors, however, had caused many problems for sales department users, and so they no longer trust the data.
- The data also use many accounting codes, which the sales department does not know how to interpret. Even though the database is now accurate, complete, consistent, and timely, sales department users may regard it as of low quality due to poor believability and interpretability.

# Major Tasks in Data Preprocessing

- Data cleaning routines work to “clean” the data by filling in missing values, smoothing noisy data, identifying or removing outliers, and resolving inconsistencies.
- If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied. Furthermore, dirty data can cause confusion for the mining procedure, resulting in unreliable output.
- Although most mining routines have some procedures for dealing with incomplete or noisy data, they are not always robust. Instead, they may concentrate on avoiding overfitting the data to the function being modeled. Therefore, a useful preprocessing step is to run your data through some data cleaning routines.

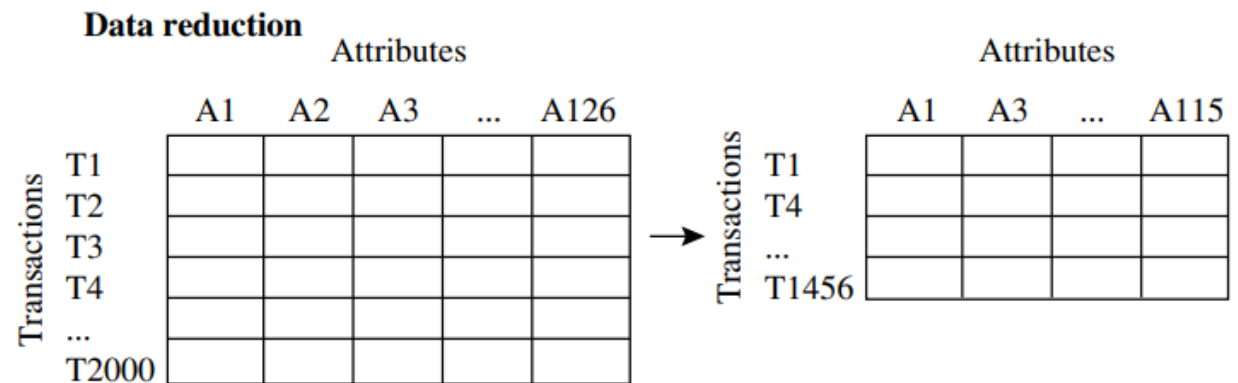
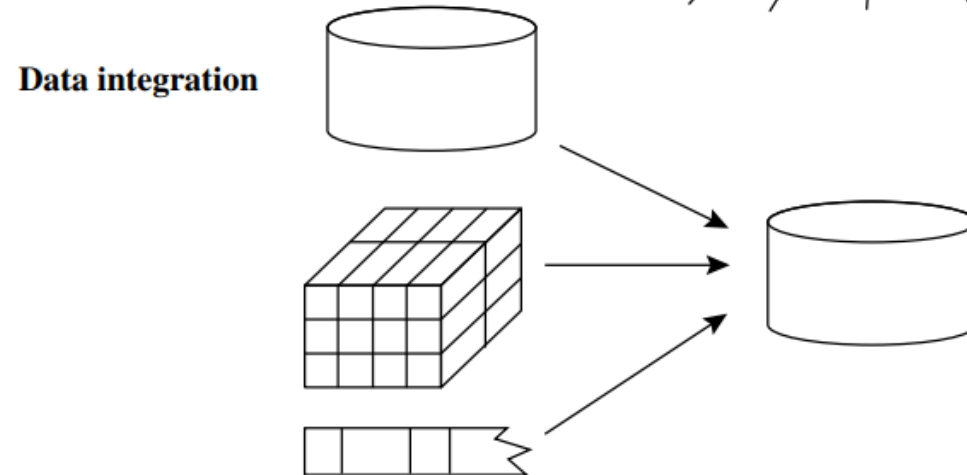
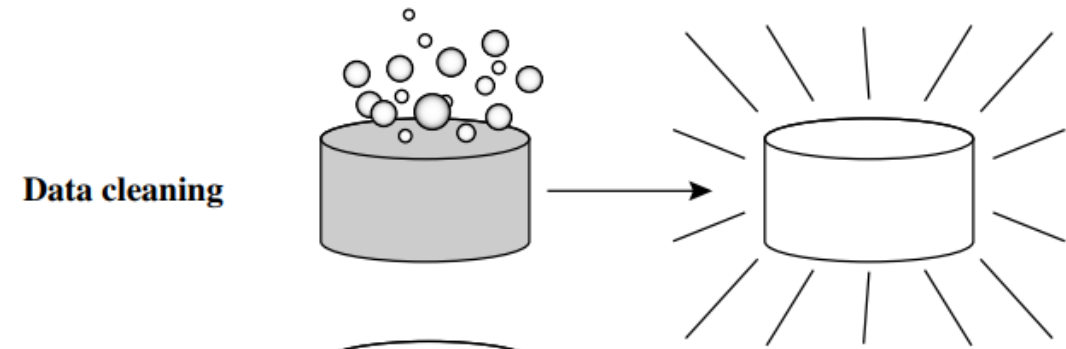
# Major Tasks in Data Preprocessing

- **Data reduction** obtains a reduced representation of the data set that is much smaller in volume, yet produces the same (or almost the same) analytical results. Data reduction strategies include dimensionality reduction and numerosity reduction.
- In **dimensionality reduction**, data encoding schemes are applied so as to obtain a reduced or “compressed” representation of the original data. Examples include **data compression techniques** (e.g., wavelet transforms and principal components analysis (PCA)), attribute subset selection (e.g., removing irrelevant attributes), and attribute construction (e.g., where a small set of more useful attributes is derived from the original set).
- In **numerosity reduction**, the data are replaced by alternative, smaller representations using parametric models (e.g., regression or log-linear models) or nonparametric models (e.g., **histograms, clusters, sampling**, or data aggregation).



# Major Tasks in Data Preprocessing

**Discretization** and **concept hierarchy generation** are powerful tools for data mining in that they allow data mining at **multiple abstraction levels**. Normalization, data discretization, and concept hierarchy generation are forms of **data transformation**.



**Data transformation**     $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

# Missing Values

- 1. Ignore the tuple:** This is usually done when the class label is missing (assuming the mining task involves classification). This method is not very effective, unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably. By ignoring the tuple, we do not make use of the remaining attributes' values in the tuple. Such data could have been useful to the task at hand.
- 2. Fill in the missing value manually:** In general, this approach is time consuming and may not be feasible given a large data set with many missing values.
- 3. Use a global constant to fill in the missing value:** Replace all missing attribute values by the same constant such as a label like "*Unknown*" or  $-\infty$ . If missing values are replaced by, say, "*Unknown*," then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common—that of "*Unknown*." Hence, although this method is simple, it is not foolproof.



# Missing Values

- 4. Use a measure of central tendency for the attribute (e.g., the mean or median) to fill in the missing value:** Chapter 2 discussed measures of central tendency, which indicate the “middle” value of a data distribution. For normal (symmetric) data distributions, the mean can be used, while skewed data distribution should employ the median (Section 2.2). For example, suppose that the data distribution regarding the income of *AllElectronics* customers is symmetric and that the mean income is \$56,000. Use this value to replace the missing value for *income*.
- 5. Use the attribute mean or median for all samples belonging to the same class as the given tuple:** For example, if classifying customers according to *credit\_risk*, we may replace the missing value with the mean *income* value for customers in the same credit risk category as that of the given tuple. If the data distribution for a given class is skewed, the median value is a better choice.
- 6. Use the most probable value to fill in the missing value:** This may be determined with regression, inference-based tools using a Bayesian formalism, or decision tree



# ETL Tools

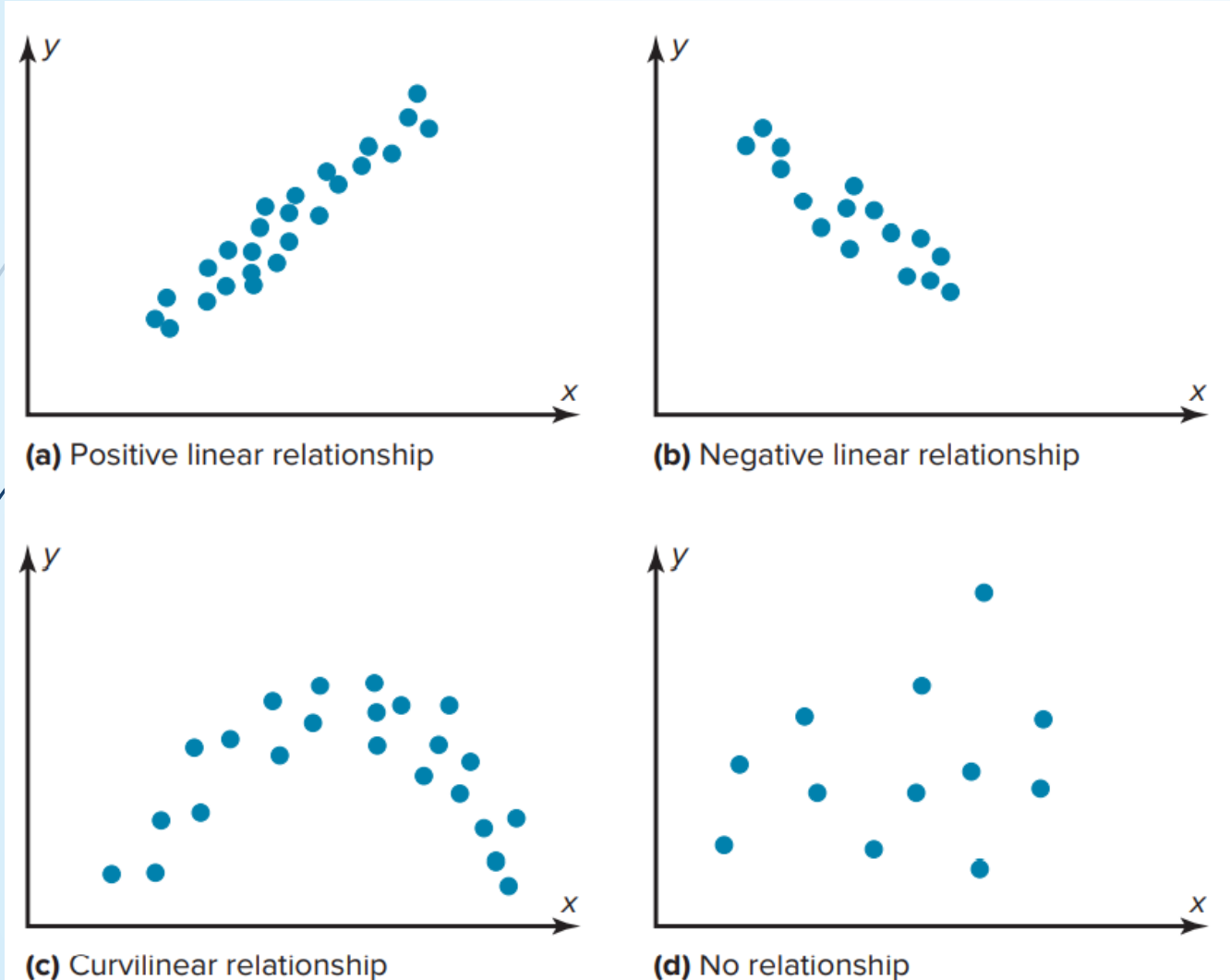
- ▶ Some data inconsistencies may be corrected manually using external references. For example, errors made at data entry may be corrected by performing a paper trace.
- ▶ Most errors, however, will require data transformations. That is, once we find discrepancies, we typically need to define and apply (a series of) transformations to correct them.
- ▶ Commercial tools can assist in the data transformation step. Data migration tools allow simple transformations to be specified **ETL (extraction/transformation/loading)** tools allow users to specify transforms through a graphical user interface (GUI). These tools typically support only a restricted set of transforms so that, often, we may also choose to write custom scripts for this step of the data cleaning process.

# Redundancy and Correlation Analysis

*Redundancy* is another important issue in data integration. An attribute (such as *annual revenue*, for instance) may be redundant if it can be “derived” from another attribute or set of attributes. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set.

Some redundancies can be detected by **correlation analysis**. Given two attributes, such analysis can measure how strongly one attribute implies the other, based on the available data. For nominal data, we use the  $\chi^2$  (*chi-square*) test. For numeric attributes, we can use the *correlation coefficient* and *covariance*, both of which access how one attribute’s values vary from those of another.

# Types of Relationships

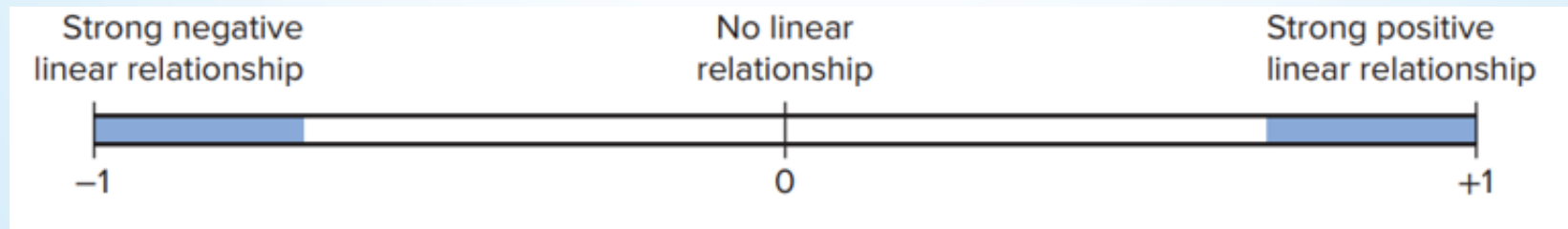


- The independent and dependent variables can be plotted on a graph called a **scatter plot**.
- The independent variable  $x$  is plotted on the horizontal axis, and the dependent variable  $y$  is plotted on the vertical axis.



# Correlation

- ▶ The range of the linear correlation coefficient is from  $-1$  to  $+1$ .
- ▶ If there is a strong positive linear relationship between the variables, the value of  $r$  will be close to  $+1$ .
- ▶ If there is a strong negative linear relationship between the variables, the value of  $r$  will be close to  $-1$ .
- ▶ When there is no linear relationship between the variables or only a weak relationship, the value of  $r$  will be close to  $0$ . See Figure 10–5. When the value of  $r$  is  $0$  or close to zero, it implies only that there is no linear relationship between the variables. The data may be related in some other **nonlinear way**.



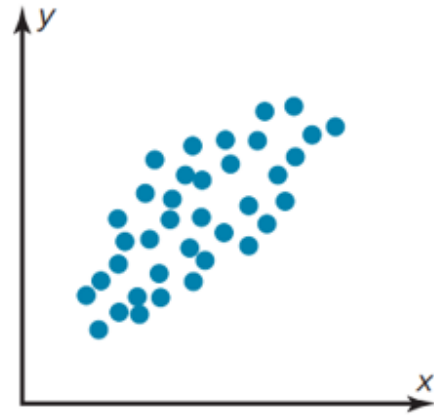


# Correlation

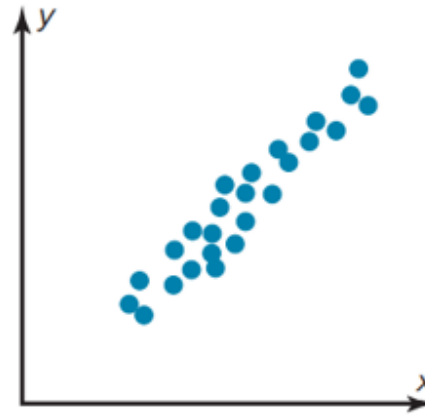
## Formula for the Linear Correlation Coefficient $r$

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}}$$

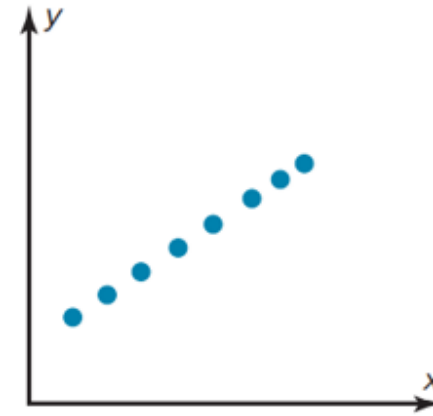
where  $n$  is the number of data pairs.



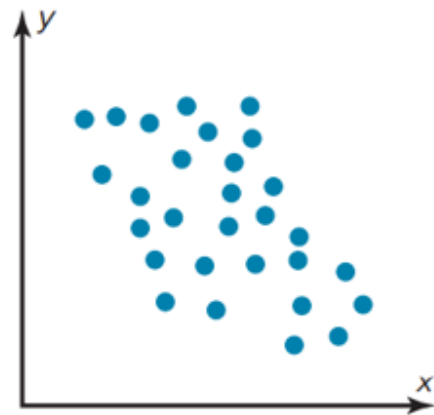
(a)  $r = 0.50$



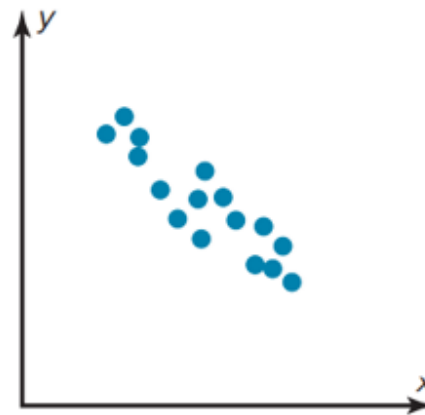
(b)  $r = 0.90$



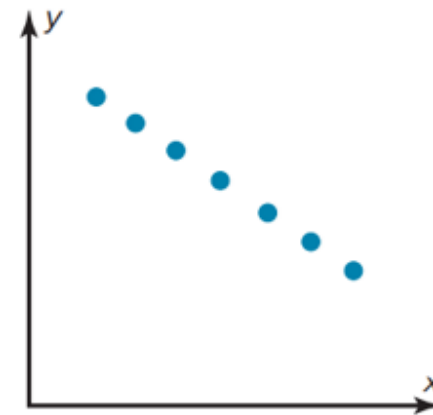
(c)  $r = 1.00$



(d)  $r = -0.50$



(e)  $r = -0.90$



(f)  $r = -1.00$

# Example

Company	Cars $x$ (in 10,000s)	Revenue $y$ (in billions of dollars)	$xy$	$x^2$	$y^2$
A	63.0	7.0	441.00	3969.00	49.00
B	29.0	3.9	113.10	841.00	15.21
C	20.8	2.1	43.68	432.64	4.41
D	19.1	2.8	53.48	364.81	7.84
E	13.4	1.4	18.76	179.56	1.96
F	8.5	1.5	12.75	72.25	2.25
	$\Sigma x = 153.8$	$\Sigma y = 18.7$	$\Sigma xy = 682.77$	$\Sigma x^2 = 5859.26$	$\Sigma y^2 = 80.67$

**Step 3** Substitute in the formula and solve for  $r$ .

$$\begin{aligned} r &= \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - (\Sigma y)^2]}} \\ &= \frac{(6)(682.77) - (153.8)(18.7)}{\sqrt{[(6)(5859.26) - (153.8)^2][(6)(80.67) - (18.7)^2]}} = 0.982 \end{aligned}$$

The linear correlation coefficient suggests a **strong positive linear relationship** between the number of cars a rental agency has and its annual revenue. That is, the more cars a rental agency has, the more annual revenue the company will have.

For nominal data, a correlation relationship between two attributes,  $A$  and  $B$ , can be discovered by a  $\chi^2$  (**chi-square**) test. Suppose  $A$  has  $c$  distinct values, namely  $a_1, a_2, \dots, a_c$ .  $B$  has  $r$  distinct values, namely  $b_1, b_2, \dots, b_r$ . The data tuples described by  $A$  and  $B$  can be shown as a **contingency table**, with the  $c$  values of  $A$  making up the columns and the  $r$  values of  $B$  making up the rows. Let  $(A_i, B_j)$  denote the joint event that attribute  $A$  takes on value  $a_i$  and attribute  $B$  takes on value  $b_j$ , that is, where  $(A = a_i, B = b_j)$ . Each and every possible  $(A_i, B_j)$  joint event has its own cell (or slot) in the table. The  $\chi^2$  value (also known as the *Pearson  $\chi^2$  statistic*) is computed as

$$\chi^2 = \sum_{i=1}^c \sum_{j=1}^r \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \quad (3.1)$$

where  $o_{ij}$  is the *observed frequency* (i.e., actual count) of the joint event  $(A_i, B_j)$  and  $e_{ij}$  is the *expected frequency* of  $(A_i, B_j)$ , which can be computed as

$$e_{ij} = \frac{\text{count}(A = a_i) \times \text{count}(B = b_j)}{n}, \quad (3.2)$$

where  $n$  is the number of data tuples,  $\text{count}(A = a_i)$  is the number of tuples having value  $a_i$  for  $A$ , and  $\text{count}(B = b_j)$  is the number of tuples having value  $b_j$  for  $B$ . The sum in Eq. (3.1) is computed over all of the  $r \times c$  cells. Note that the cells that contribute the most to the  $\chi^2$  value are those for which the actual count is very different from that expected.

The  $\chi^2$  statistic tests the hypothesis that  $A$  and  $B$  are *independent*, that is, there is no correlation between them. The test is based on a significance level, with  $(r - 1) \times (c - 1)$  degrees of freedom. We illustrate the use of this statistic in Example 3.1. If the hypothesis can be rejected, then we say that  $A$  and  $B$  are statistically correlated.

# Redundancy and Correlation Analysis

**Table 3.1** Example 2.1's  $2 \times 2$  Contingency Table Data

	<i>male</i>	<i>female</i>	<i>Total</i>
<i>fiction</i>	250 (90)	200 (360)	450
<i>non-fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

*Note:* Are *gender* and *preferred\_reading* correlated?

# Redundancy and Correlation Analysis

**Correlation analysis of nominal attributes using  $\chi^2$ .** Suppose that a group of 1500 people was surveyed. The gender of each person was noted. Each person was polled as to whether his or her preferred type of reading material was fiction or nonfiction. Thus, we have two attributes, *gender* and *preferred\_reading*. The observed frequency (or count) of each possible joint event is summarized in the contingency table shown in Table 3.1, where the numbers in parentheses are the expected frequencies. The expected frequencies are calculated based on the data distribution for both attributes using Eq. (3.2).

Using Eq. (3.2), we can verify the expected frequencies for each cell. For example, the expected frequency for the cell (*male*, *fiction*) is

$$e_{11} = \frac{\text{count}(\text{male}) \times \text{count}(\text{fiction})}{n} = \frac{300 \times 450}{1500} = 90,$$

and so on. Notice that in any row, the sum of the expected frequencies must equal the total observed frequency for that row, and the sum of the expected frequencies in any column must also equal the total observed frequency for that column.



# Redundancy and Correlation Analysis

**Table 3.1** Example 2.1's  $2 \times 2$  Contingency Table Data

	<i>male</i>	<i>female</i>	<i>Total</i>
<i>fiction</i>	250 (90)	200 (360)	450
<i>non_fiction</i>	50 (210)	1000 (840)	1050
Total	300	1200	1500

*Note:* Are *gender* and *preferred\_reading* correlated?

Using Eq. (3.1) for  $\chi^2$  computation, we get

$$\begin{aligned}\chi^2 &= \frac{(250 - 90)^2}{90} + \frac{(50 - 210)^2}{210} + \frac{(200 - 360)^2}{360} + \frac{(1000 - 840)^2}{840} \\ &= 284.44 + 121.90 + 71.11 + 30.48 = 507.93.\end{aligned}$$

For this  $2 \times 2$  table, the degrees of freedom are  $(2 - 1)(2 - 1) = 1$ . For 1 degree of freedom, the  $\chi^2$  value needed to reject the hypothesis at the 0.001 significance level is 10.828 (taken from the table of upper percentage points of the  $\chi^2$  distribution, typically available from any textbook on statistics). Since our computed value is above this, we can reject the hypothesis that *gender* and *preferred\_reading* are independent and conclude that the two attributes are (strongly) correlated for the given group of people. ■

# Covariance of Numeric Data

In probability theory and statistics, correlation and covariance are two similar measures for assessing how much two attributes change together. Consider two numeric attributes  $A$  and  $B$ , and a set of  $n$  observations  $\{(a_1, b_1), \dots, (a_n, b_n)\}$ . The mean values of  $A$  and  $B$ , respectively, are also known as the **expected values** on  $A$  and  $B$ , that is,

$$E(A) = \bar{A} = \frac{\sum_{i=1}^n a_i}{n}$$

and

$$E(B) = \bar{B} = \frac{\sum_{i=1}^n b_i}{n}.$$

The **covariance** between  $A$  and  $B$  is defined as

$$\text{Cov}(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}. \quad (3.4)$$

If we compare Eq. (3.3) for  $r_{A,B}$  (correlation coefficient) with Eq. (3.4) for covariance, we see that

$$r_{A,B} = \frac{\text{Cov}(A, B)}{\sigma_A \sigma_B}, \quad (3.5)$$

where  $\sigma_A$  and  $\sigma_B$  are the standard deviations of  $A$  and  $B$ , respectively. It can also be shown that

$$\text{Cov}(A, B) = E(A \cdot B) - \bar{A}\bar{B}. \quad (3.6)$$

This equation may simplify calculations.

# Covariance of Numeric Data

**Table 3.2** Stock Prices for *AllElectronics* and *HighTech*

<i>Time point</i>	<i>AllElectronics</i>	<i>HighTech</i>
t1	6	20
t2	5	10
t3	4	14
t4	3	5
t5	2	5

**Covariance analysis of numeric attributes.** Consider Table 3.2, which presents a simplified example of stock prices observed at five time points for *AllElectronics* and *HighTech*, a high-tech company. If the stocks are affected by the same industry trends, will their prices rise or fall together?

$$E(\text{AllElectronics}) = \frac{6 + 5 + 4 + 3 + 2}{5} = \frac{20}{5} = \$4$$

and

$$E(\text{HighTech}) = \frac{20 + 10 + 14 + 5 + 5}{5} = \frac{54}{5} = \$10.80.$$

Thus, using Eq. (3.4), we compute

$$\begin{aligned} \text{Cov}(\text{AllElectronics}, \text{HighTech}) &= \frac{6 \times 20 + 5 \times 10 + 4 \times 14 + 3 \times 5 + 2 \times 5}{5} - 4 \times 10.80 \\ &= 50.2 - 43.2 = 7. \end{aligned}$$

Therefore, given the positive covariance we can say that stock prices for both companies rise together. ■

*Variance* is a special case of covariance, where the two attributes are identical (i.e., the covariance of an attribute with itself). Variance was discussed in Chapter 2.




# Data Reduction

- **Dimensionality reduction** Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration.
- **Numerosity reduction** techniques replace the original data volume by alternative, smaller forms of data representation.
- In **data compression**, transformations are applied so as to obtain a reduced or “compressed” representation of the original data. If the original data can be reconstructed from the compressed data without any information loss, the data reduction is called lossless. If, instead, we can reconstruct only an approximation of the original data, then the data reduction is called lossy





# Dimensionality Reduction

- ▶ Wavelet Transform
  - ▶ Principal Component Analysis (PCA)
  - ▶ Attribute Subset Selection
- 



# Data Reduction

## Histograms

- ▶ 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.

# Data Reduction

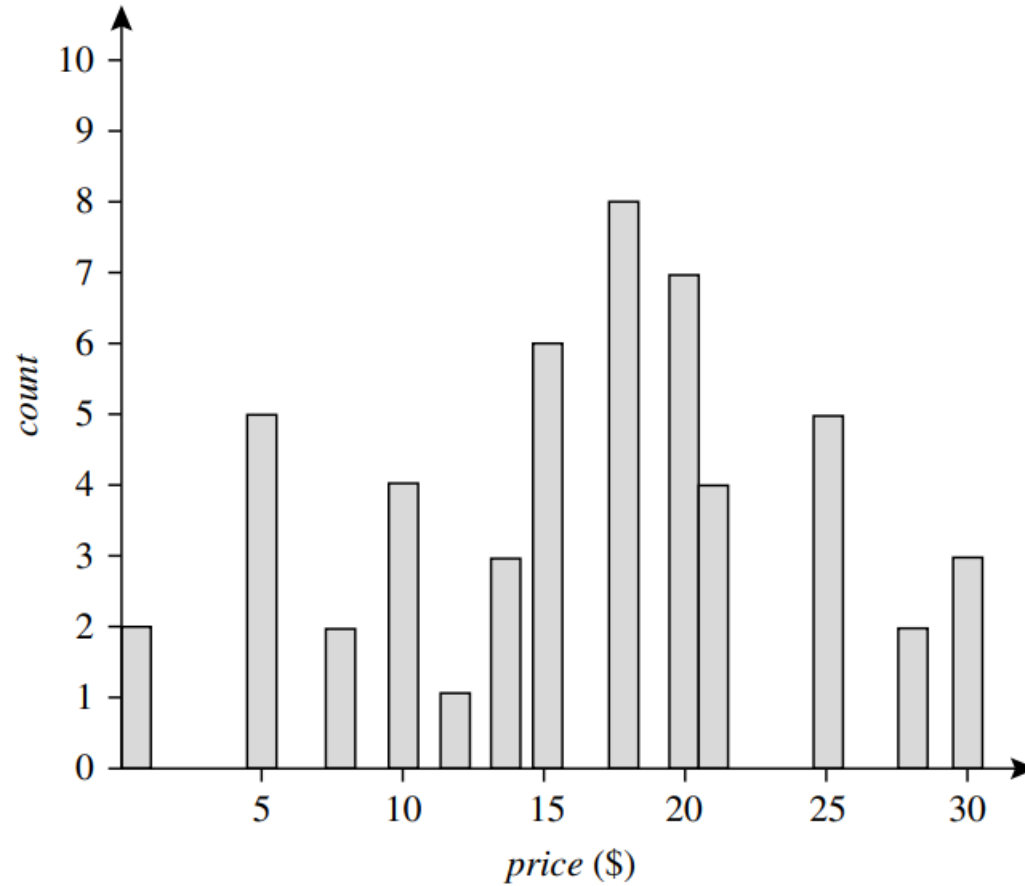
## Histograms

**Histograms.** The following data are a list of *AllElectronics* prices for commonly sold items (rounded to the nearest dollar). 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, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Figure 3.7 shows a histogram for the data using singleton buckets. To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute. In Figure 3.8, each bucket represents a different \$10 range for *price*. ■

# Data Reduction

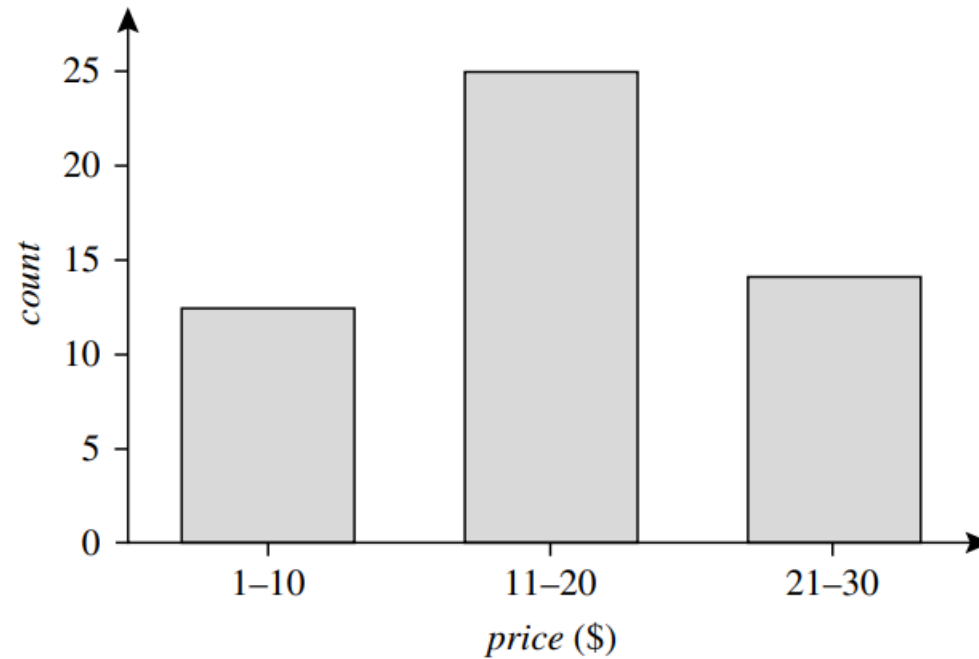
## Histograms



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

# Data Reduction

## Histograms



---

An equal-width histogram for *price*, where values are aggregated so that each bucket has a uniform width of \$10.





# Data Reduction

## Histograms

*“How are the buckets determined and the attribute values partitioned?”* There are several partitioning rules, including the following:

- **Equal-width:** In an equal-width histogram, the width of each bucket range is uniform (e.g., the width of \$10 for the buckets in Figure 3.8).
- **Equal-frequency** (or equal-depth): In an equal-frequency histogram, the buckets are created so that, roughly, the frequency of each bucket is constant (i.e., each bucket contains roughly the same number of contiguous data samples).



# Sampling

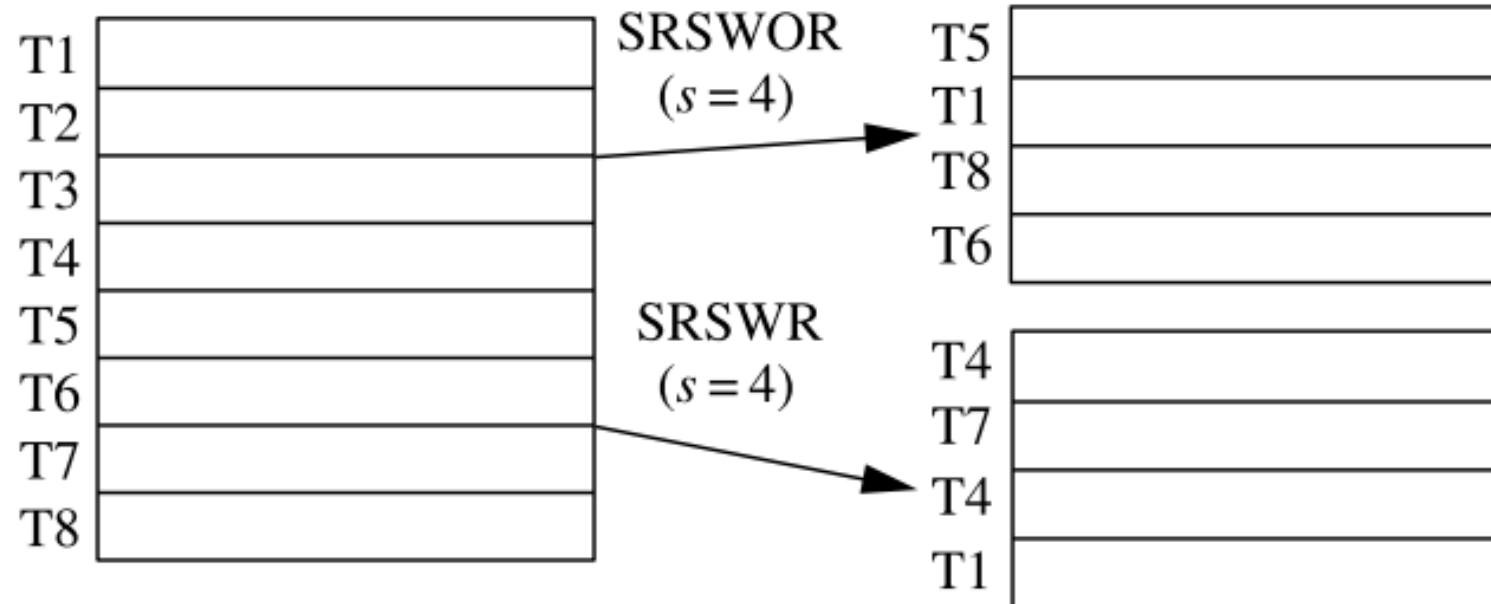
- ▶ Sampling can be used as a data reduction technique because it allows a large data set to be represented by a much smaller random data sample (or subset).
- ▶ Suppose that a large data set,  $D$ , contains  $N$  tuples. Let's look at the most common ways that we could sample  $D$  for data reduction, as illustrated in the next figure:



# Sampling

- **Simple random sample without replacement (SRSWOR)** of size  $s$ : This is created by drawing  $s$  of the  $N$  tuples from  $D$  ( $s < N$ ), where the probability of drawing any tuple in  $D$  is  $1/N$ , that is, all tuples are equally likely to be sampled.
- **Simple random sample with replacement (SRSWR)** of size  $s$ : This is similar to SRSWOR, except that each time a tuple is drawn from  $D$ , it is recorded and then replaced. That is, after a tuple is drawn, it is placed back in  $D$  so that it may be drawn again.

# Sampling

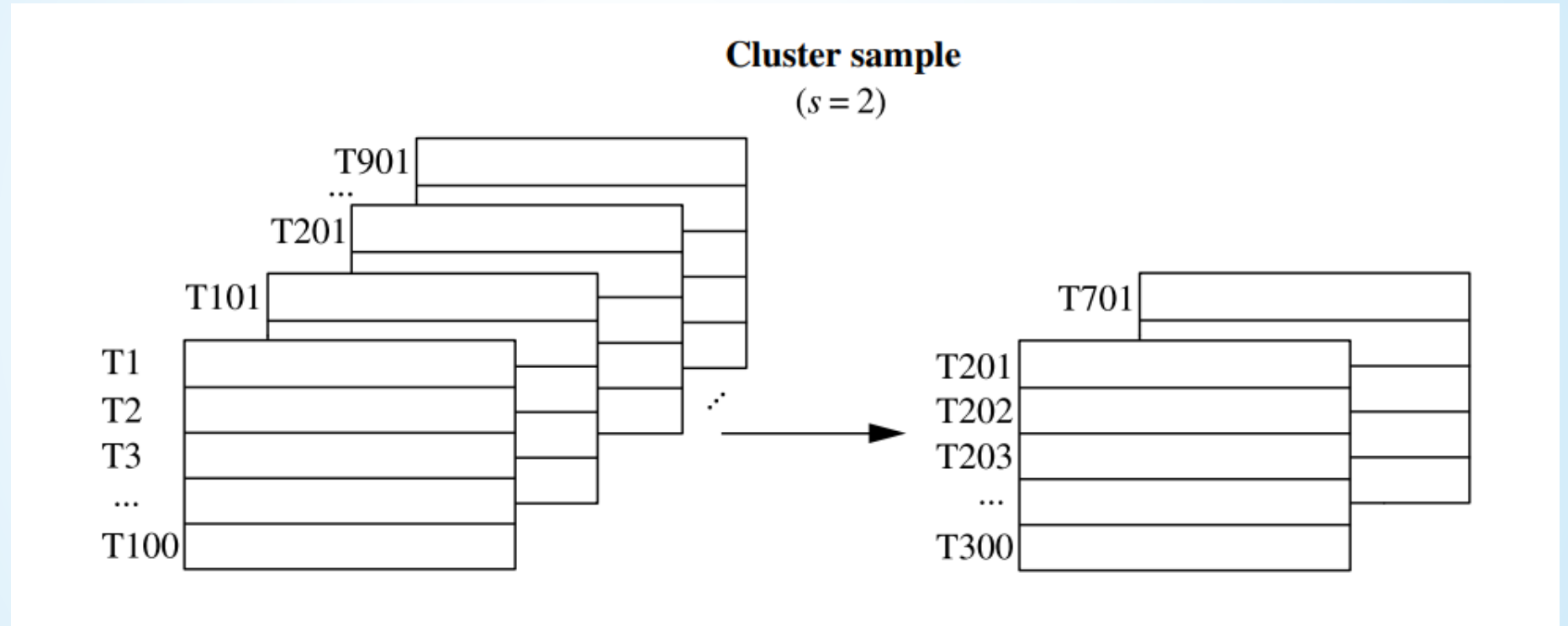


# Sampling

- **Cluster sample:** If the tuples in  $D$  are grouped into  $M$  mutually disjoint “clusters. For example, tuples in a database are usually retrieved a page at a time, so that each page can be considered a cluster. A reduced data representation can be obtained by applying, say, SRSWOR to the pages, resulting in a cluster sample of the tuples. Other clustering criteria conveying rich semantics can also be explored. For example, in a spatial database, we may choose to define clusters geographically based on how closely different areas are located.
- **Stratified sample:** If  $D$  is divided into mutually disjoint parts called strata, a stratified sample of  $D$  is generated by obtaining an SRS at each stratum. This helps ensure a representative sample, especially when the data are skewed. For example, a stratified sample may be obtained from customer data, where a stratum is created for each customer age group. In this way, the age group having the smallest number of customers will be sure to be represented.



# Sampling



# Sampling

## Stratified sample (according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

# Reference

Data Mining, Concepts and Techniques,  
Jiawei Han, Micheline Kamber, Jian Pei.  
MK. Chapter 3.

