Data, Measurements, and Data Preprocessing

Table of Contents

1 Introduction & Learning Objectives

Data Quality Matters

Garbage In, Garbage Out

The quality of data directly impacts the quality of data mining results; flawed data leads to flawed outcomes. Therefore, data quality is important.

Understanding Data Nature

Before applying any mining algorithm, it is crucial to understand the characteristics and nature of the data being used. This ensures appropriate methods are applied.

Learning Objectives

01 Classifying Data Types

Learn to classify data into different attribute types: nominal, binary, ordinal, and numeric data.

02 Distinguishing Data Types

Learn the differences between discrete and continuous data and their implications for analysis.

03 Computing Basic Statistical Measures

Learn to compute and interpret basic statistical measures for effective data description, leading to better insights.

04 Visualizing Data Statistics

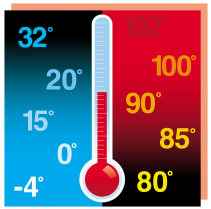
Gain the ability to visualize basic data statistics for enhanced understanding and communication of key data features.

2 Data Types (Attribute Types)

Attribute Definition

Attribute Explained

An attribute (or feature/variable) is a data field representing a characteristic of a data object, such as temperature, color, or ID. Attributes are fundamental.



Nominal Attributes: Categories Without Order

Operations on Nominal Attributes

Only equality (=) and inequality (≠) operations are meaningful for nominal attributes; mode is the only central tendency measure. No arithmetic operations apply.

Definition of Nominal Attributes

Nominal attributes are categories, names, or labels with no inherent order; examples include colors, cities, or employee IDs.

Meaningless Averages

Note that calculating an "average city" or comparing IDs (ID1 < ID2) does not make sense for nominal data. Labels are not comparable.

Example 2.1. Nominal attributes. Suppose that *hair_color* and *marital_status* are two attributes describing *person* objects. In our application, possible values for *hair_color* are *black*, *brown*, *blond*, *red*, *auburn*, *gray*, and *white*. The attribute *marital_status* can take on the values *single, married, divorced*,

and *widowed*. Both *hair_color* and *marital_status* are nominal attributes. Another example of a nominal attribute is *occupation*, with the values *teacher, dentist, programmer, farmer*, and so on. □

Binary Attributes: Two Categories

Operations on Binary Attributes

Similar to nominal attributes, but special similarity measures like the Jaccard coefficient exists for asymmetric binary attributes. Similarity calculation expands possibilities.

Definition of Binary Attributes

Binary attributes are a special case of nominal attributes with only two categories, such as gender (Male, Female) or a medical test result (Positive, Negative).

Symmetric vs. Asymmetric Binary Attributes

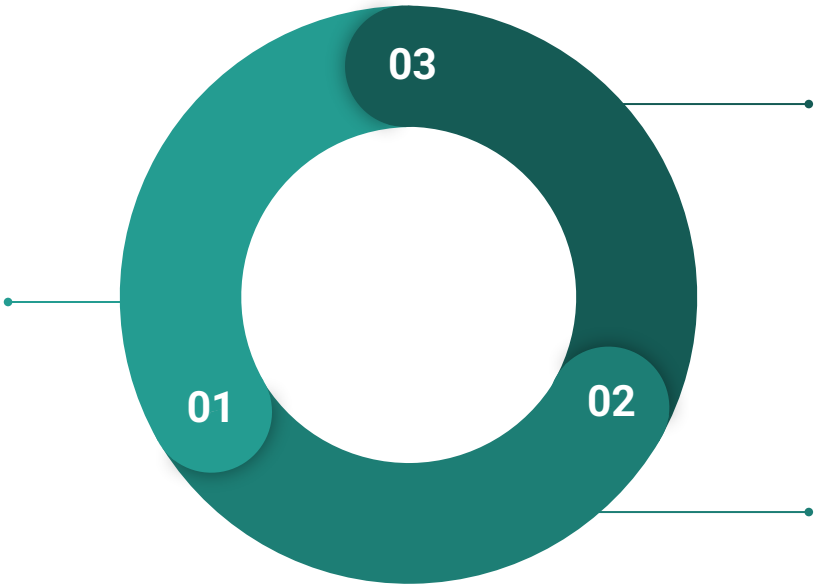
Types of Binary Attributes: Symmetric states are equally valuable, whereas asymmetric attributes have one state that is more significant or rare.

Example 2.2. Binary attributes. Given the attribute *smoker* describing a *patient* object, 1 indicates that the patient smokes, whereas 0 indicates that the patient does not. Similarly, suppose the patient undergoes a medical test that has two possible outcomes. The attribute *medical_test* is binary, where a value of 1 means the result of the test for the patient is positive, whereas 0 means the result is negative. □

Ordinal Attributes: Meaningful Order

Definition of Ordinal Attributes

Values have a meaningful order or ranking, but differences between values are not quantifiable, e.g., education level or satisfaction rating. Prioritize ordering for analysis.



Key Point: Quantifying Differences

While we know Excellent > Good, we cannot say how much better it is numerically for ordinal data; quantify difference for more insights.

Operations on Ordinal Attributes

Comparison operators (<, >) are meaningful; median and percentile are valid measures of central tendency. Comparisons are valid for ordered data.

Example 2.3. Ordinal attributes. 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*.

The values have a meaningful sequence (which corresponds to increasing drink size); however, we cannot tell from the values *how much* bigger, say, a large is than a medium. Other examples of ordinal attributes include *grade* (e.g., $A+$, A, $A-$, $B+$, and so on) and *professional_rank*. Professional ranks can be enumerated in a sequential order: for example, *assistant*, *associate*, and *full* for professors, and *private, private second class, private first class, specialist, corporal, sergeant, ...* for army ranks.

Ordinal attributes are useful for registering subjective assessments of qualities that cannot be measured objectively; thus ordinal attributes are often used in surveys for ratings. In one survey, participants were asked to rate how satisfied they were as customers. Customer satisfaction had the following ordinal categories: *1: very dissatisfied, 2: dissatisfied, 3: neutral, 4: satisfied*, and *5: very satisfied.* □

Numeric Attributes: Quantitative Data

Interval-Scaled Attributes

Interval-scaled attributes have meaningful differences, but no true zero point, e.g., temperature in °C or °F. Ratios are not meaningful.

Ratio-Scaled Attributes

Ratio-scaled attributes have a true zero point; ratios are meaningful, examples include height, weight, and age; all arithmetic operations are valid. Calculate valid ratios.

Definition of Numeric Attributes

Numeric attributes are quantitative; values are real numbers where arithmetic operations make sense. Calculation operations are valid.

Discrete vs. Continuous Attributes

Discrete attributes have a finite or countably infinite set of values, while continuous attributes are real numbers with potentially infinite values within a range. Analyze value set.

Example 2.4. Interval-scaled attributes. A *temperature* attribute is interval-scaled. Suppose that we have the outdoor *temperature* values 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*. In addition, we can quantify the difference between values. For example, a temperature of 20 °C is five degrees higher than a temperature of 15 °C. Calendar dates are another example. For instance, the years 2012 and 2020 are eight years apart. □

Example 2.5. Ratio-scaled attributes. Unlike temperatures in Celsius and Fahrenheit, the Kelvin (K) temperature scale has what is considered a true zero-point (0 K = −273.15 °C): It is the point at which all thermal motion ceases in the classical description of thermodynamics. Other 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). Additional examples include attributes to measure weight, height, and speed, and monetary quantities (e.g., you are 100 times richer with $100 than with $1). □

3 Statistics of Data (Descriptive Statistics)

Purpose of Statistics: Understand Data Behavior

Understand General Behavior

Summarize and understand the general behavior of data before mining to gain valuable insights and prepare data for effective analysis. Describe behavioral indicators.

Measuring Central Tendency

Mode

The most frequent value(s), applicable to all attribute types (nominal, ordinal, numeric); a dataset can be unimodal, bimodal, or multimodal. Report all frequency measures.

Mean (Average)

Applicable to numeric data but sensitive to outliers; for example, the mean salary in a company can be skewed by one CEO's high salary. Exclude or adjust outliers.

Median

The middle value when data is sorted, robust to outliers, applicable to ordinal and numeric data; for an even number of observations, it is the average of two middle values. Compute robust measure.

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_N}{N}. \tag{2.1}$$

This corresponds to the built-in aggregate function, *average* (`avg()` in SQL), provided in relational database systems.

Example 2.6. Mean. Suppose we have the following values for *salary* (in thousands of dollars), shown in ascending order: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Using Eq. (2.1), we have

$$\begin{aligned}\bar{x} &= \frac{30 + 36 + 47 + 50 + 52 + 52 + 56 + 60 + 63 + 70 + 70 + 110}{12} \\ &= \frac{696}{12} = 58.\end{aligned}$$

Thus, the mean salary is $58,000. □

$$\bar{x} = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i} = \frac{w_1 x_1 + w_2 x_2 + \cdots + w_N x_N}{w_1 + w_2 + \cdots + w_N}.$$

This is called the **weighted arithmetic mean** or the **weighted average**.

Example 2.7. Median. Let's find the median of the data from Example 2.6. The data are already sorted in ascending order. There is an even number of observations (i.e., 12); therefore, the median is not unique. It can be any value within the two middlemost values of 52 and 56 (that is, within the sixth and seventh values in the list). By convention, we assign the average of the two middlemost values as the median; that is, $\frac{52+56}{2} = \frac{108}{2} = 54$. Thus, the median is \$54,000.

Suppose that we had only the first 11 values in the list. Given an odd number of values, the median is the middlemost value. This is the sixth value in this list, which has a value of \$52,000. □

$$median \approx L_1 + \left(\frac{N/2 - \left(\sum freq\right)_l}{freq_{median}} \right) \times width, \tag{2.3}$$

where L_1 is the lower boundary of the median interval, N is the number of values in the entire data set, $\left(\sum freq\right)_l$ is the sum of the frequencies of all of the intervals that are lower than the median interval, $freq_{median}$ is the frequency of the median interval, and *width* is the width of the median interval.

Example 2.8. Mode. The data from Example 2.6 are bimodal. The two modes are \$52,000 and \$70,000. □

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

$$mean - mode \approx 3 \times (mean - median). \tag{2.4}$$

Measuring Dispersion (Spread)

Range

Max - Min; is simple but highly sensitive to outliers, thus the IQR is more typically used. Consider IQR as a potential countermeasure.

Variance s^2 and Standard Deviation s

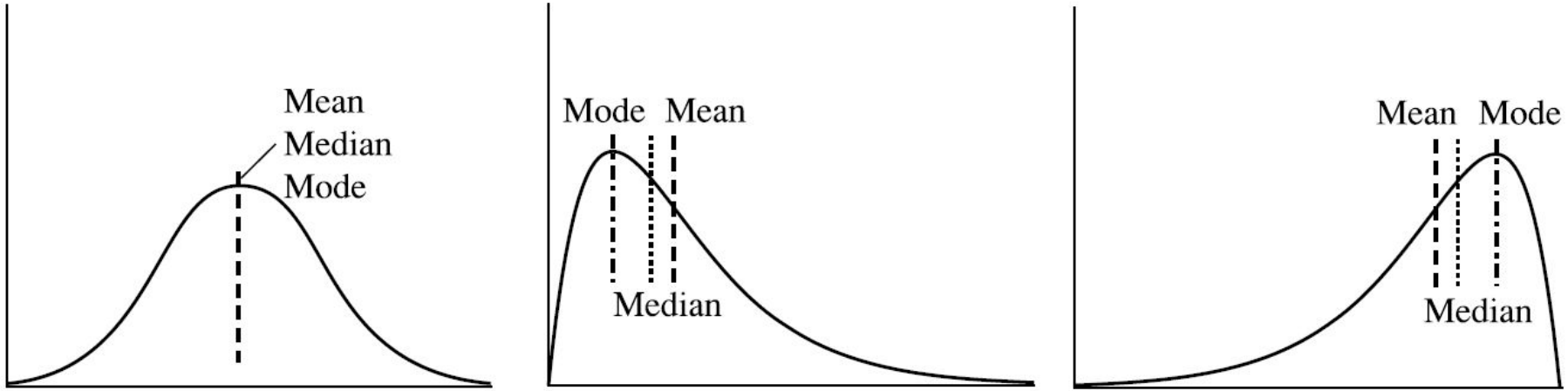
Measure the average squared deviation from the mean; standard deviation is in the same units as the data and easier to interpret; low deviations means data points are close to the mean. Calculate average squared deviation.

Interquartile Range (IQR)

IQR = Q3 - Q1, where Q1 is the 25th percentile and Q3 is the 75th percentile. The range of the middle 50% of the data. Robust to outliers. Used in box plots. Filter out outliers.

Example 2.9. Midrange. The midrange of the data of Example 2.6 is $\frac{30{,}000+110{,}000}{2} = \$70{,}000.$ □

In a unimodal frequency curve with perfect **symmetric** data distribution, the mean, median, and mode are all at the same center value, as shown in Fig. 2.1a.



(a) Symmetric data **(b)** Positively skewed data **(c)** Negatively skewed data

FIGURE 2.1

Mean, median, and mode of symmetric vs. positively and negatively skewed data.

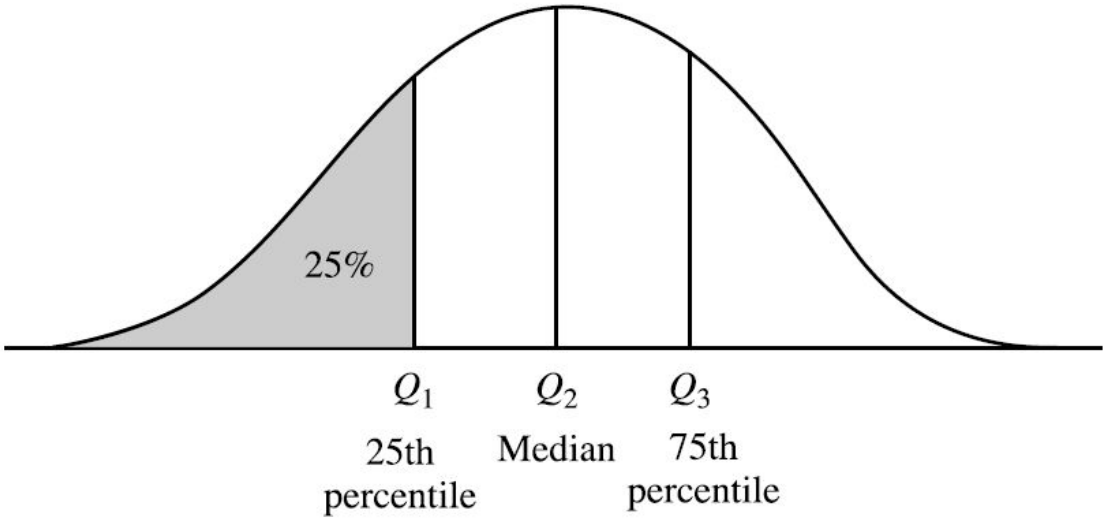


FIGURE 2.2

A plot of the data distribution for some attribute X. The quantiles plotted are quartiles. The three quartiles divide the distribution into four equal-size consecutive subsets. The second quartile corresponds to the median.

$$IQR = Q_3 - Q_1. \tag{2.5}$$

Example 2.10. Interquartile range. The quartiles are the three values that split the sorted data set into four equal parts. The data of Example 2.6 contain 12 observations, already sorted in ascending order. Since there are even number of elements on this list, the median of the list should be the mean of the center two elements, that is ($52,000 + $56,000)/2 = $54,000. Then the first quartile should be the mean of the 3rd and 4th elements, that is, ($47,000 + $50,000)/2 = $48,500, whereas the 3rd quartile should be the mean of the 9th and 10th elements, that is, ($63,000 + $70,000)/2 = $66,500. Thus the interquartile range is $IQR = \$66{,}500 - \$48{,}500 = \$18{,}000$. □

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{N}x_i^2\right) - \bar{x}^2, \tag{2.6}$$

where $\bar{x}$ is the mean value of the observations, as defined in Eq. (2.1). The **standard deviation**, σ, of the observations is the square root of the variance, σ^2.

Order Data: 8, 9, 10, 12, 14, 15

Position of $Q_1 = 0.25 \times (n + 1) = 0.25 \times 7 = 1.75$

$Q1 = 8 + 0.75 \times (9 - 8) = 8.75$

Position of $Q_2 = 0.5 \times (n + 1) = 0.5 \times 7 = 3.5$

$Q2 = 10 + 0.5 \times (12 - 10) = 11$

Position of $Q_3 = 0.75 \times (n + 1) = 0.75 \times 7 = 5.25$

$Q3 = 14 + 0.25 \times (15 - 14) = 14.25$

$IQR = Q3 - Q1 = 14.25 - 8.75 = 5.5$ cm

Example 2.12. Variance and standard deviation. In Example 2.6, we found $\bar{x} = \$58{,}000$ using Eq. (2.1) for the mean. To determine the variance and standard deviation of the data from that example, we set $N = 12$ and use Eq. (2.6) to obtain

$$\begin{aligned}\sigma^2 &= \frac{1}{12}(30^2 + 36^2 + 47^2 \ldots + 110^2) - 58^2 \\ &\approx 379.17 \\ \sigma &\approx \sqrt{379.17} \approx 19.47.\end{aligned}$$

□

Covariance and Correlation Analysis

Covariance

Measures how two numeric variables change together; the sign indicates direction: positive = variables move together, negative = they move inversely; magnitude is scale-dependent. Assess variables together.

Correlation Coefficient (Pearson's r)

Normalized covariance, ranging from -1 to +1; +1 is a perfect positive linear relationship, 0 is no linear relationship, -1 is a perfect negative linear relationship; does not imply causation. Correlation is not Causation.

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

and

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

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

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

Mathematically, it can also be shown that

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

Table 2.1 Stock prices for *AllElectronics* and *HighTech*.

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

Example 2.13. Covariance analysis of numeric attributes. Consider Table 2.1, 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?

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

and

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

Thus, using Eq. (2.7), we compute

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

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

$$r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A\sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\bar{A}\bar{B}}{n\sigma_A\sigma_B}, \tag{2.9}$$

where n is the number of tuples, a_i and b_i are the respective values of A and B in tuple i, $\bar{A}$ and $\bar{B}$ are the respective mean values of A and B, σ_A and σ_B are the respective standard deviations of A and B (as defined in Section 2.2.2), and $\Sigma(a_i b_i)$ is the sum of the AB cross-product (i.e., for each tuple, the value for A is multiplied by the value for B in that tuple). Note that $-1 \leq r_{A,B} \leq +1$. If $r_{A,B}$ is greater than 0, then A and B are *positively correlated*, meaning that the values of A increase as the values of B increase. The higher the value, the stronger the correlation (i.e., the more each attribute implies the other). Hence, a higher value may indicate that A (or B) may be removed as a redundancy.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}, \tag{2.10}$$

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

$$e_{ij} = \frac{count(A = a_i) \times count(B = b_j)}{n}, \tag{2.11}$$

where n is the number of data tuples, $count(A = a_i)$ is the number of tuples having value a_i for A, and $count(B = b_j)$ is the number of tuples having value b_j for B. The sum in Eq. (2.10) is computed over all of the $r \times c$ cells. Note that the cells that contribute the most to the χ^2 value are those for which the actual count is very different from that expected.

Example 2.14. Correlation analysis of nominal attributes using χ^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 2.2, where the numbers in parentheses are the expected frequencies. The expected frequencies are calculated based on the data distribution for both attributes using Eq. (2.11).

Using Eq. (2.11), we can verify the expected frequencies for each cell. For example, the expected frequency for the cell (*male, fiction*) is

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

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.

Using Eq. (2.10) for χ^2 computation, we get

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

For this 2×2 table, the degrees of freedom are $(2-1) \times (2-1) = 1$. For 1 degree of freedom, the χ^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 χ^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. □

Table 2.2 Example 2.1's 2 x 2 contingency table data.

	Male	Female	Total
fiction	250 (90)	200 (360)	450
non_fiction	50 (210)	1000 (840)	1050
Total	300	1200	1500

Note: Are gender and preferred_reading correlated?

Graphic Displays of Basic Statistics

01 Box Plot (Box-and-Whisker Plot)

Visually displays: Min, Q1, Median, Q3, Max, and Outliers; excellent for comparing distributions across groups. Visualize and compare data.

02 Scatter Plot

For two numeric variables, revealing relationships, clusters, outliers for enhanced data interpretation. Expose underlying relationships.

03 Histogram

For numeric data, visualizing frequency distribution and revealing skewness and modality. Reveal distributions.

04 Bar Chart / Pie Chart

For categorical data (nominal/ordinal) charting to show distributions. Clearly show distributions.

Table 2.3 A set of unit price data for items sold at a branch of the online store.

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

Example 2.15. Quantile plot. Fig. 2.4 shows a quantile plot for the *unit price* data of Table 2.3.

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

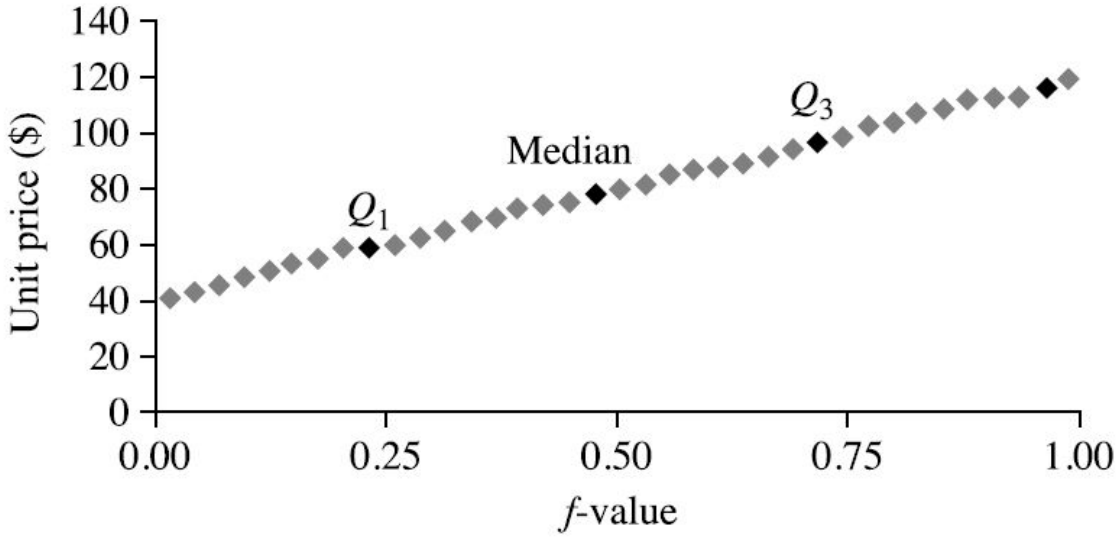


FIGURE 2.4

A quantile plot for the unit price data of Table 2.3.

Example 2.16. Quantile-quantile plot. Fig. 2.5 shows a quantile-quantile plot for *unit price* data of items sold at two branches of the online store during a given time period. Each point corresponds to the same quantile for each data set and shows the unit price of items sold at branch 1 vs. branch 2 for that quantile. (To aid comparison, the straight line represents the case where, for each given quantile, the unit price at each branch is the same. The darker points correspond to the data for Q_1, the median, and Q_3, respectively.)

We see, for example, that at Q_1, the unit price of items sold at branch 1 was slightly less than that at branch 2. In other words, 25% of items sold at branch 1 were less than or equal to \$60, whereas 25% of items sold at branch 2 were less than or equal to \$64. At the 50th percentile (marked by the median, which is also Q_2), we see that 50% of items sold at branch 1 were less than \$78, whereas 50% of items at branch 2 were less than \$85. In general, we note that there is a shift in the distribution of branch 1 with respect to branch 2 in that the unit prices of items sold at branch 1 tend to be lower than those at branch 2. □

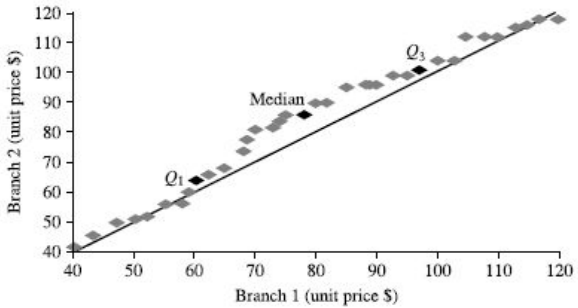


FIGURE 2.5

A q-q plot for unit price data from two branches of the online store.

Example 2.17. Histogram. Fig. 2.6 shows a histogram for a data set on research award distribution for a region, where buckets (or bins) are defined by equal-width ranges representing \$1000 increments, and the frequency is the number of research awards in the corresponding buckets. □

Although histograms are widely used, they may not be as effective as the quantile plot, q-q plot, and boxplot methods in comparing groups of univariate observations.

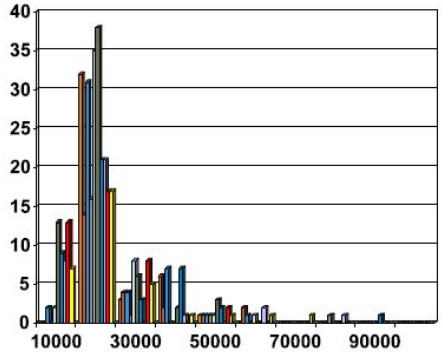


FIGURE 2.6

A histogram on research award distribution for a region.

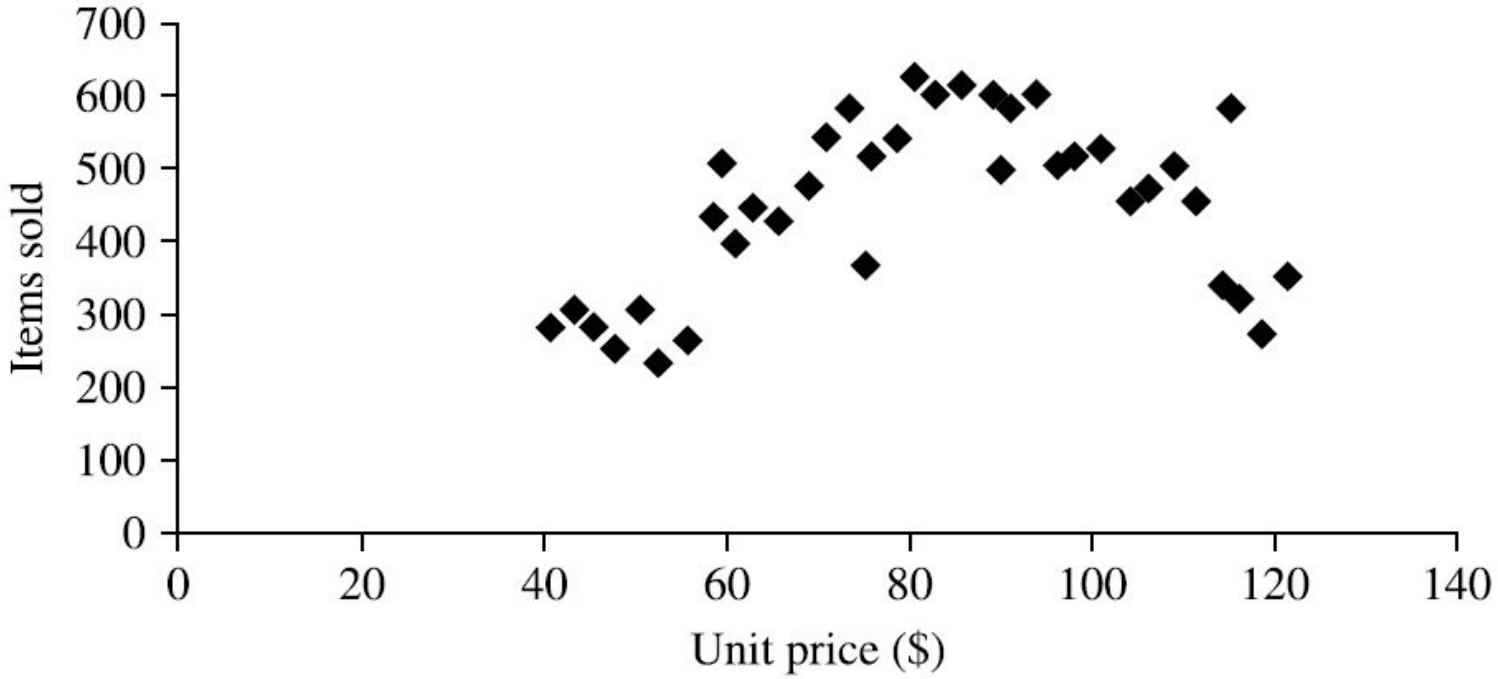


FIGURE 2.7

A scatter plot for Table 2.3 data set.

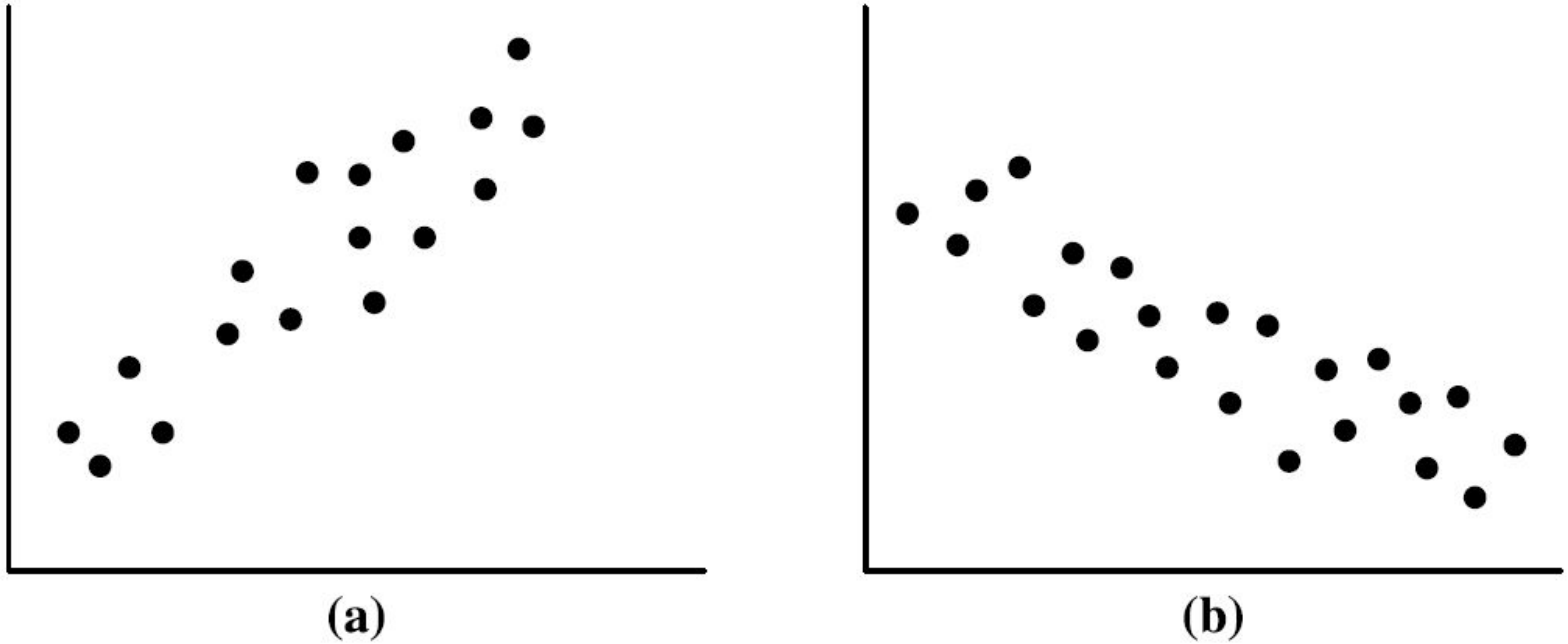


FIGURE 2.8

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

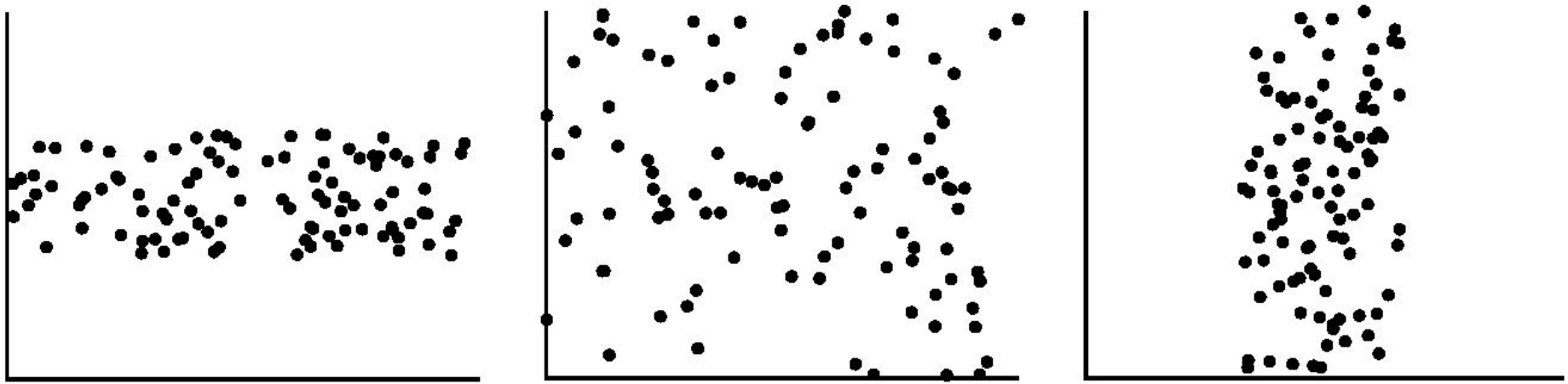
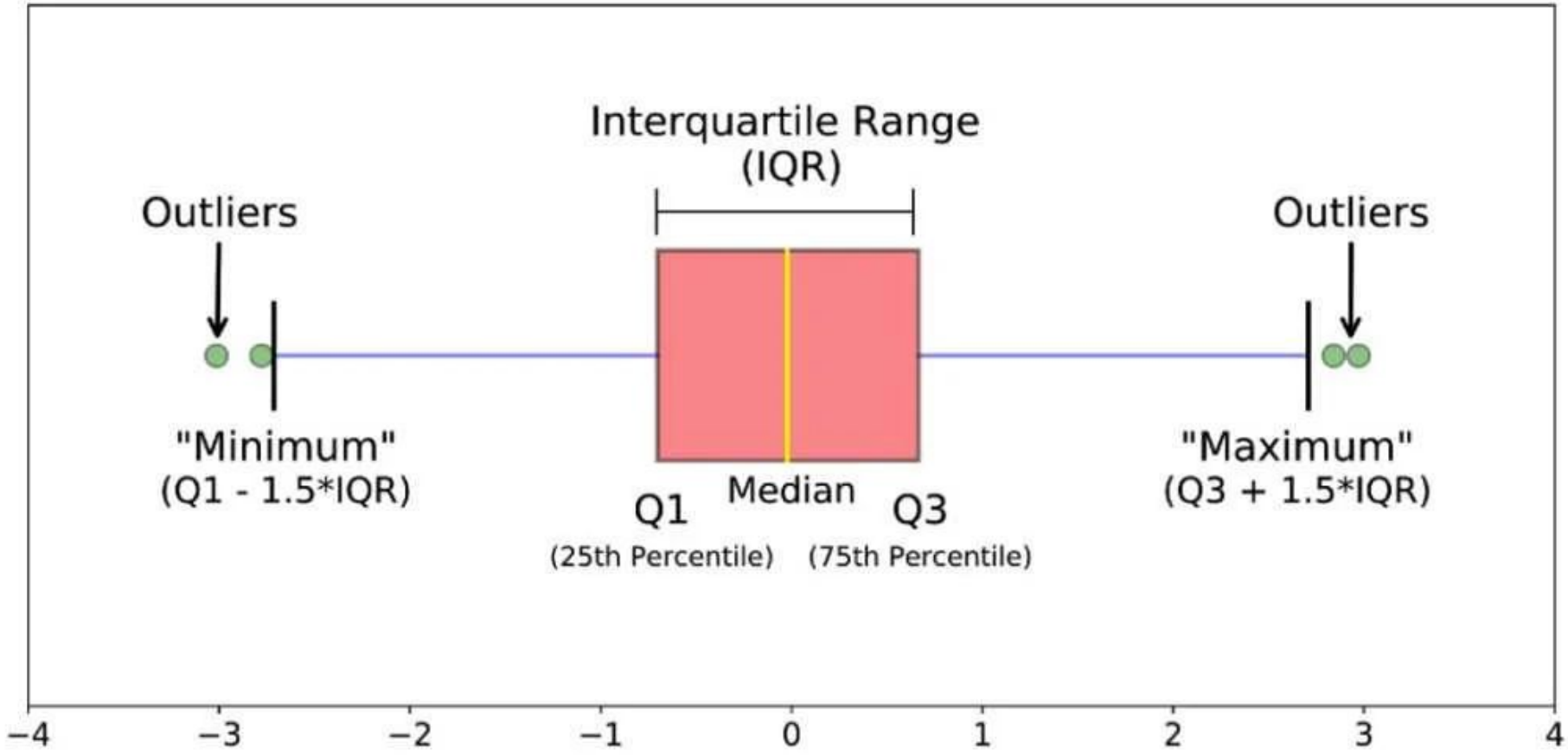


FIGURE 2.9

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


Interquartile Range
(IQR)
Outliers
Outliers
"Minimum"
(Q1 - 1.5*IQR)
"Maximum"
(Q3 + 1.5*IQR)
Q1
Median
Q3
(25th Percentile)
(75th Percentile)
−4
−3
−2
−1
0
1
2
3
4

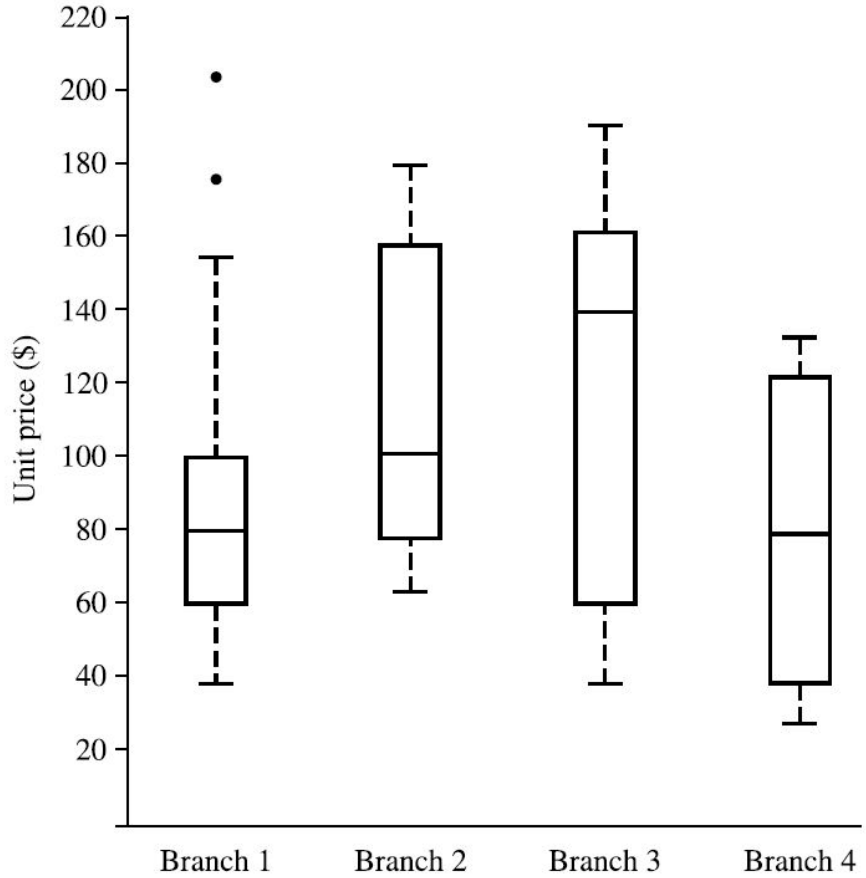


FIGURE 2.3

Boxplot for the unit price data for items sold at four branches of an online store during a given time period.

4 Summary, Implications for Data Mining & Preview

Why This Matters for Data Mining

Algorithm Choice

K-means requires numeric data; decision trees can handle nominal data; choose the appropriate algorithm based on data type. Consider data when choosing.

Data Preprocessing

Choose correct normalization; you must know the data type to choose correct normalization (e.g., min-max for numeric, one-hot encoding for nominal). Normalize based on data type.

Error Detection

Understanding statistics helps spot anomalies (e.g., a negative age) before mining; detect and correct errors early for data accuracy. Identify potentially incorrect data.