1. **What is meant by attribute? Explain the different types of attributes with examples.**

Attributes (describe objects)., Variable, field, characteristic, feature or observation

Attribute Type	Description	Examples	Operations
Nominal	Each value represents a label. (Typical comparisons between two values are limited to "equal" or "no equal")	Flower color, gender, zip code	Mode, entropy, contingency correlation, χ^2 test
Ordinal	The values can be ordered. (Typical comparisons between two values are "equal" or "greater" or "less")	Hardness of minerals, {good, better, best}, grades, street numbers, rank, age	Median, percentiles, rank correlation, run tests, sign tests
Interval	The differences between values are meaningful, i.e., a unit of measurement exists. (+, -)	Calendar dates, temperature in Celsius or Fahrenheit	Mean, standard deviation, Pearson's correlation, t and F tests
Ratio	Differences and ratios are meaningful. (*, /)	Monetary quantities, counts, age, mass, length, electrical current	Geometric mean, harmonic mean, percent variation

Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object

Nominal: categories, states, or names

Hair_color = {auburn, black, blond, brown, grey, red, white}

marital status, occupation, ID numbers, zip codes

Binary: Nominal attribute with only 2 states (0 and 1)

Symmetric binary: both outcomes equally important

e.g., gender

Asymmetric binary: outcomes not equally important

e.g., medical test (positive vs. negative)

Convention: assign 1 to most important outcome (e.g., HIV positive)

Ordinal

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

Size = {small, medium, large}, grades, army rankings

Quantity (integer or real-valued)

Interval
Measured on a scale of equal-sized units Values have order
E.g., temperature in C˚or F˚, calendar dates
No true zero-point
Ratio
Inherent zero-point., We can speak of values as being an order of magnitude larger than the unit of measurement (10 K˚ is twice as high as 5 K˚)
e.g., temperature in Kelvin, length, counts,
monetary quantities.

2. Explain the measures of central tendency and dispersion.

- Mean (algebraic measure) (sample vs. population): $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\mu = \frac{\sum x}{N}$
  Note: *n* is sample size and *N* is population size
  - Weighted arithmetic mean:
  - Trimmed mean: chopping extreme values
- Median: $\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
  - Middle value if odd number of values, or average of the middle two values otherwise
  - Estimated by interpolation (for *grouped data*):

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

age	*frequency*
1–5	200
6–15	450
16–20	300
21–50	1500
51–80	700
81–110	44

- Mode
  - Value that occurs most frequently in the data
  - Unimodal, bimodal, trimodal
  - Empirical formula:

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

Measuring the Dispersion of Data

- Quartiles, outliers and boxplots
  - **Quartiles**: Q_1 (25th percentile), Q_3 (75th percentile)
  - **Inter-quartile range**: IQR = $Q_3 - Q_1$
  - **Five number summary**: min, Q_1, median, Q_3, max
  - **Boxplot**: ends of the box are the quartiles; median is marked; add whiskers, and plot outliers individually
  - **Outlier**: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ*)
  - **Variance**: (algebraic, scalable computation)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2] \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2 = \frac{1}{N}\sum_{i=1}^{n} x_i^2 - \mu^2$$

  - **Standard deviation** *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

3. Explain the concept of master data management.

The processes and technology to produce and maintain a single clean copy of master data → The "Golden" record

An Application for creating and maintaining an authoritative view of master data including policies and procedures for access, update, modification, viewing between systems across the enterprise

- MDM is the process of helping a company to standardize the definition and attributes of all of its critical data elements (customer, vendor, product, etc.) to create a common point of reference enterprise wide.
- MDM can facilitate the sharing of data among all a company's disparate business functions, departments and even divisions - not to mention across all information systems, platforms and applications

4. List the major issues in data mining.

- Mining different kinds of knowledge in databases
- Interactive mining of knowledge at multiple levels of abstraction
- Incorporation of background knowledge
- Data mining query languages and ad hoc data mining
- Presentation and visualization of data mining results
- Handling noisy or incomplete data
- Efficiency and scalability of data mining algorithms
- Pattern evaluation
- Handling of relational and complex types of data
- Parallel, distributed, and incremental mining algorithms
- Mining information from heterogeneous databases and global information systems

5. What is meant by data cleaning? Explain the basic methods for data cleaning

How to Handle Noisy Data?

- Binning
  - first sort data and partition into (equal-frequency) bins
  - then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.
- Regression
  - smooth by fitting the data into regression functions
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values and check by human (e.g., deal with possible outliers)

Data in the Real World Is Dirty: Lots of potentially incorrect data, e.g., instrument faulty, human or computer error, transmission error

- incomplete: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
  - e.g., *Occupation*=" " (missing data)
- noisy: containing noise, errors, or outliers
  - e.g., *Salary*="−10" (an error)
- inconsistent: containing discrepancies in codes or names, e.g.,
  - *Age*="42", *Birthday*="03/07/2010"
  - Was rating "1, 2, 3", now rating "A, B, C"
  - discrepancy between duplicate records
- Intentional (e.g., *disguised missing* data)
  - Jan. 1 as everyone's birthday?

6. Explain data placement and query parallelism with reference to parallel databases

Physical placement of the DB onto multiple nodes

Static vs. Dynamic