



Dr. Vishwanath Karad

**MIT WORLD PEACE
UNIVERSITY** | PUNE

TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Data Science by

Shilpa Sonawani

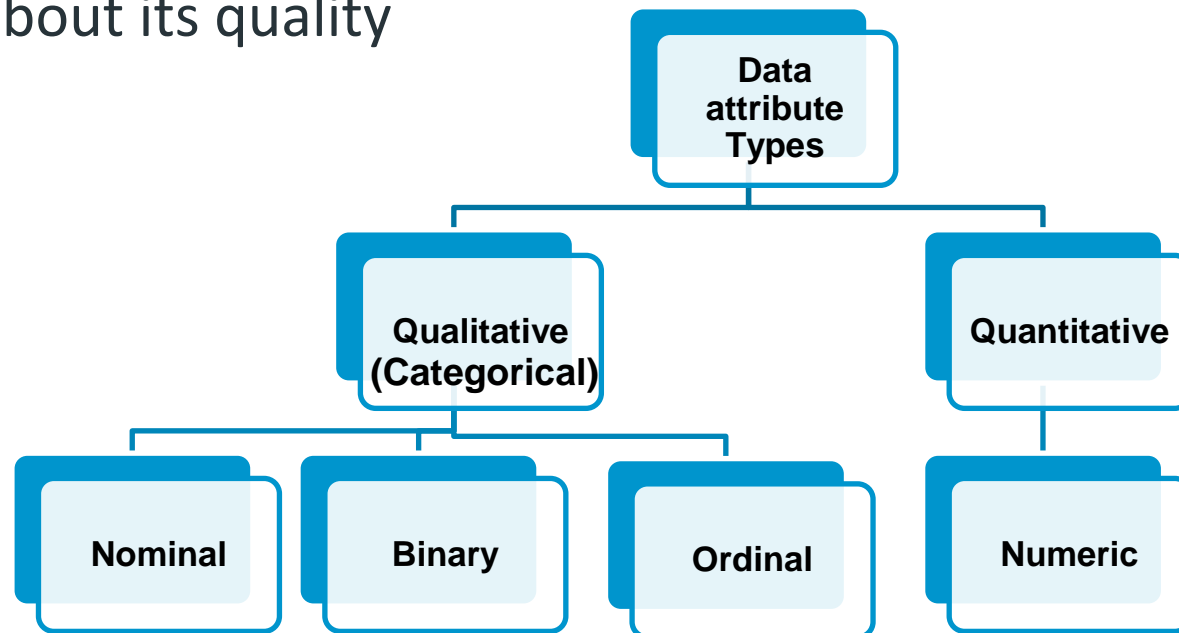
SCHOOL OF COMPUTER ENGINEERING AND TECHNOLOGY

Why Is Data Preprocessing Important?

- No quality data, no quality mining results!
 - **Quality** decisions must be based on quality data
 - e.g., duplicate or missing data may cause incorrect or even misleading statistics.
- Data in the real world is dirty
 - **Incomplete**: e.g., Occupation=""
 - **Noisy**: e.g., Salary="-10"
 - **Inconsistent**: e.g., Age="42" Birthday="03/07/1997"
e.g., Was rating "1,2,3", now rating "A, B, C"
e.g., Discrepancy between duplicate records
- Data cleaning and transformation comprise the majority of the work in a data mining application (90%).

Understanding Data Attribute Types

- For a customer, attributes can be customer-id, address, etc
- First step before Data pre-processing - differentiate the data between different types of attributes
 - Quantitative: You can count (Say, Quantity – has some numerical value)
 - Qualitative: You can not count and has no numeric value, but you can think about its quality



Qualitative Attributes

1. Nominal Attributes –

- Values are name of things or some kind of symbols (without quantitative/numeric value)
- Values of nominal attributes represents some category or state and that's why nominal attribute also referred as **categorical attributes** and
- There is no order (rank, position) among values of nominal attribute

Attribute	Values
Colours	Black, Brown, White

Other Examples: Hair Color, Nationalities, Names of People, and so on

Qualitative Attributes

2. Binary Attributes : Binary data has only 2 values/states.

For Example yes or no, true or false.

- i. **Symmetric** : Both values are equally important (Gender). No preference on which should be coded as 0 or 1
- ii. **Asymmetric** : Both values are not equally important. Most important outcome is coded as 1

Attribute	Values
Gender	Male , Female

Attribute	Values
Cancer detected	Yes, No
result	Pass , Fail

Qualitative Attributes

3. Ordinal Attributes :

- Values that have a meaningful sequence or ranking (order) between them
- But magnitude of values is not actually known

ORDINAL DATA

Ordinal data classifies variables into categories which have a natural order or rank.

Examples

School grades



How is ordinal data analyzed?

Education level



Descriptive statistics:
Frequency distribution, mode, median, and range

Seniority level



Non-parametric statistical tests

Quantitative Attributes

Numeric :

- A numeric attribute is quantitative because, it is a measurable quantity, represented as integer or real values.
- Numerical attributes are of 2 types, **interval** and **ratio**.

i) Interval-scaled attributes

- It is concerned with both the order and difference between your variables. This allows you to measure standard deviation and central tendency.
- Values can be added and subtracted but cannot be multiplied or divided
- e.g. Temperature of 20°C is warmer than 10°C , and the difference between 20 deg and 10 deg is 10°C . The difference between 10 deg and 0 deg is also 10°C.

Quantitative Attributes(Contd..)

- Interval data always appears in the form of numbers or numerical values where the distance between the two points is **standardized** and **equal**
- Do not have a true zero even if one of the values carries the name “zero”
 - A true zero has no value, but 0 degrees C definitely has a value: it's quite chilly. You can also have negative numbers.
 - If you don't have a true zero, you can't calculate ratios. This means addition and subtraction work, but division and multiplication don't.

ii) Ratio-scaled attributes

- Has all properties of interval-scaled
- Have a true zero
 - A good example of ratio data is weight in kilograms. If something weighs zero kilograms, it truly weighs nothing—compared to temperature (interval data), where a value of zero degrees doesn't mean there is “no temperature,” it simply means it's extremely cold!
- Values can be added , subtracted , multiplied & divided
- e.g. Weight, height, etc

Quantitative Attributes(Contd..)

- Interval variables, ratio variables can be discrete or continuous.
- A **discrete variable** is expressed only in countable numbers (e.g., integers)
- A **continuous variable** can potentially take on an infinite number of values.

	The four levels of measurement			
	Nominal	Ordinal	Interval	Ratio
Categories	✓	✓	✓	✓
Rank order		✓	✓	✓
Equal spacing			✓	✓
True zero				✓

Exercise

- Classify the following attributes as discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Nominal: Categories without order

Ordinal: Categories with meaningful order

Interval: No 'True' zero. Division makes no sense

Ratio: 'True' zero. Exits. Division makes sense

1. Number of telephones in your house
2. Size of French Fries (Medium or Large or X-Large)
3. Ownership of a cell phone
4. Number of local phone calls you made in a month
5. Length of longest phone call
6. Length of your foot
7. Price of your textbook
8. Zip code
9. Temperature in degrees Fahrenheit
10. Temperature in degrees Celsius
11. Temperature in Kelvin

Major Tasks in Data Preprocessing

- **Data Cleaning**
 - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- **Data Integration**
 - Integration of multiple databases, or files
- **Data Transformation**
 - Normalization and aggregation
- **Data Reduction**
 - Obtains reduced representation in volume but produces the same or similar analytical results
- **Data Discretization (for numerical data)**

Data Cleaning Tasks

- Importance
Data cleaning is the first step
- Fill in missing values
- Identify outliers and smooth out noisy data
- Correct inconsistent data
- Resolve redundancy caused by data integration



Data Cleaning: Incomplete (Missing) Data

- Data is not always available
 - Many tuples may not have recorded value for several attributes, such as customer income in sales data
- Missing data may be due to
 - Equipment malfunction
 - May not be available at the time of entry
 - Data not entered due to misunderstanding
 - Certain data may not be considered important at the time of entry
 - Data inconsistent with other recorded data may have been deleted
 - Recording of data or its modifications may have been overlooked
- Information is not collected
 - (e.g., people decline to give their age and weight)
 - Attributes may not be applicable to all cases (e.g., annual income is not applicable to children)
- Missing data may need to be inferred

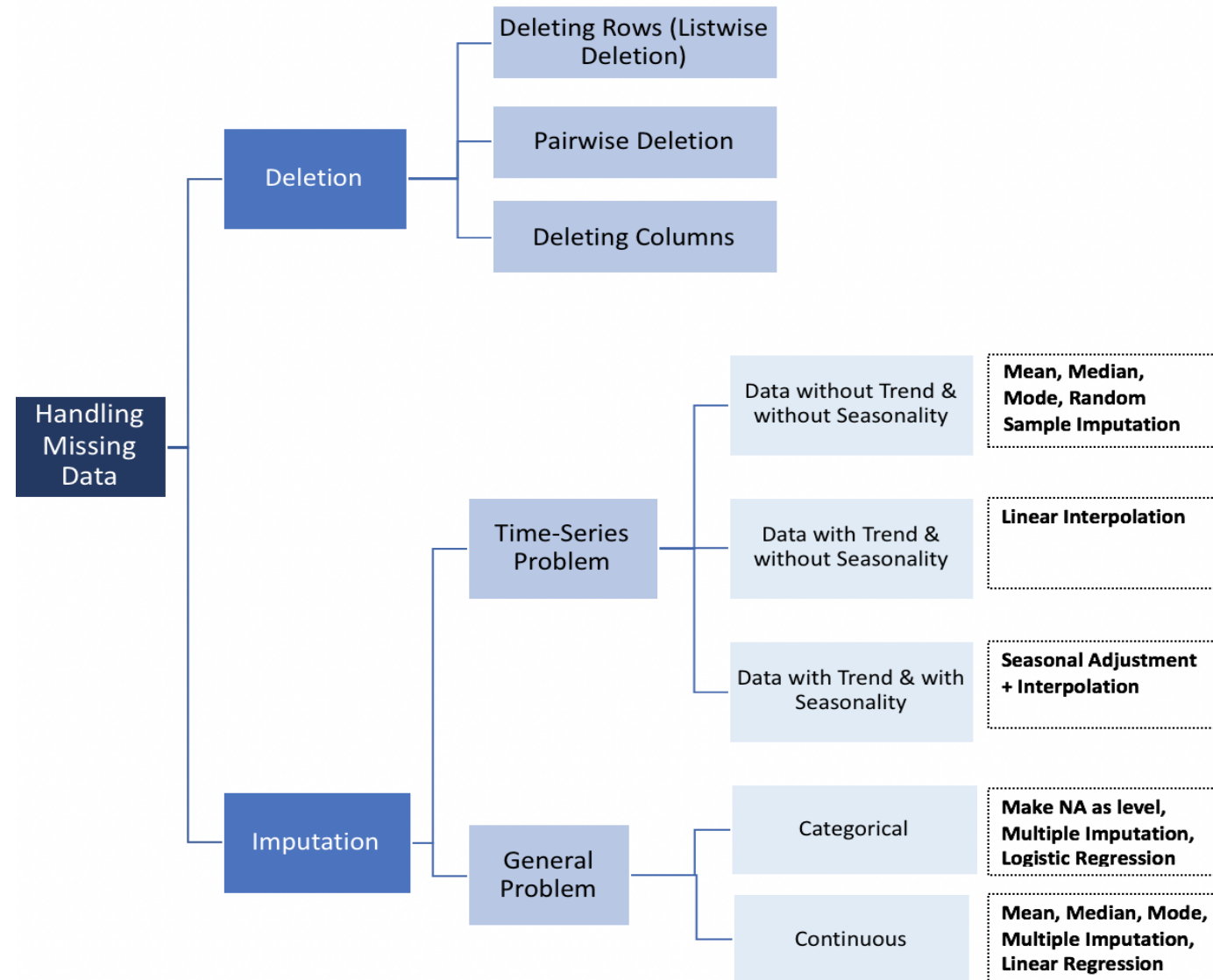
Python Collections to Store the Data

- Python **Collections** are used to store data.
- Four types: *lists, tuple, dictionaries, sets, and tuples*, all of which are built-in collections
 - **List** is a collection which is ordered and changeable. Allows duplicate members.
e.g. `[1, 2, 3, 4, 5]`
 - **Tuple** is a collection which is ordered and unchangeable. Allows duplicate members.
e.g. `(1, 2, 3, 4, 5)`
 - **Set** is a collection which is unordered, unchangeable and unindexed. No duplicate members. But, you can remove and/or add items whenever you like.
e.g. `{1, 2, 3, 4, 5}`
 - **Dictionary** is a collection which is ordered (from Python version 3.7) and changeable. No duplicate members.
e.g. `{1: "a", 2: "b", 3: "c", 4: "d", 5: "e"}`

How to Handle Missing Data?

- Ignore the tuple: not effective unless tuple contains several attributes with missing values
- Fill in the missing value manually: tedious + infeasible?
- Fill it automatically with
 - A global constant
 - The attribute mean or median
 - The attribute mean for all samples belonging to the same class as the tuple: smarter
 - The most probable value: inference-based such as Bayesian formula or decision tree

How to Handle Missing Data?



Noisy Data

- **Noise**: random error or variance in a measured variable
- **Incorrect attribute values** may be due to
 - Faulty data collection instruments
 - Data entry problems
 - Data transmission problems
 - Technology limitation
 - Inconsistency in naming convention
- **Other data problems** which require data cleaning
 - Duplicate records
 - Incomplete data
 - Inconsistent data

How to Handle Noisy Data?

- Binning
 - First sort data and partition into (equal-frequency) bins
 - Then one can smooth by bin mean, 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 manually

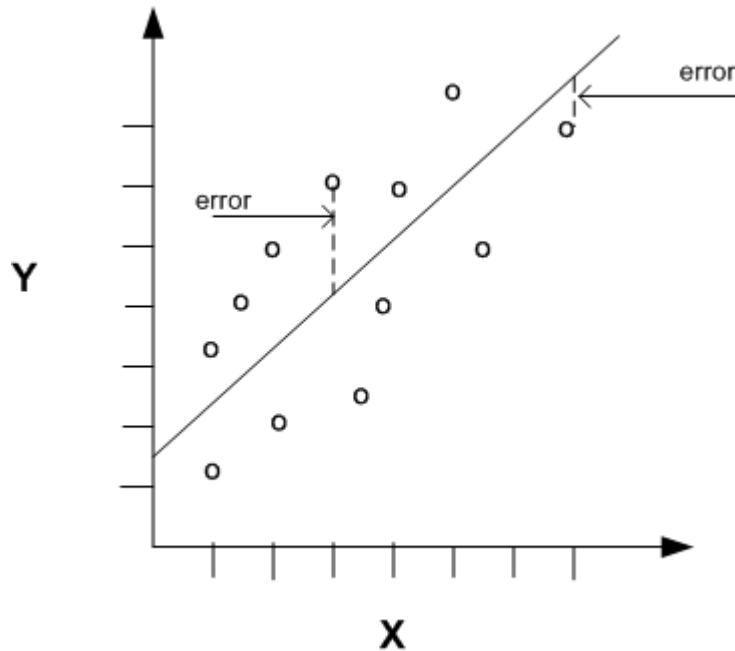
Binning Methods for Data Smoothing

- Sorted data for price (in dollars): 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34
- Partition into (equi-depth) bins:
 - Bin 1: 4, 8, 9, 15
 - Bin 2: 21, 21, 24, 25
 - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
 - Bin 1: 9, 9, 9, 9
 - Bin 2: 23, 23, 23, 23
 - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries:
 - Bin 1: 4, 4, 15, 15 (Replaced by closest boundary)
 - Bin 2: 21, 21, 25, 25
 - Bin 3: 26, 26, 34, 34

Exercise

- Suppose a group of 12 sales price records has been sorted as follows:
5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 Partition them into three bins solve it by each of the following methods:
 - (a) equi-depth partitioning
 - (b) Smoothing by bin boundaries
- Solution:
 - (a) equi-depth partitioning -Bin-1: 5, 10, 11, 13, Bin-2: 15, 35, 50, 55, Bin-03: 72, 92, 204, 215
 - (b) Smoothing by bin boundaries: Bin-1: 5,13,13,13 , Bin-2: 15,15,55,55, Bin-3:72,72,215,215

Regression for Data Smoothing



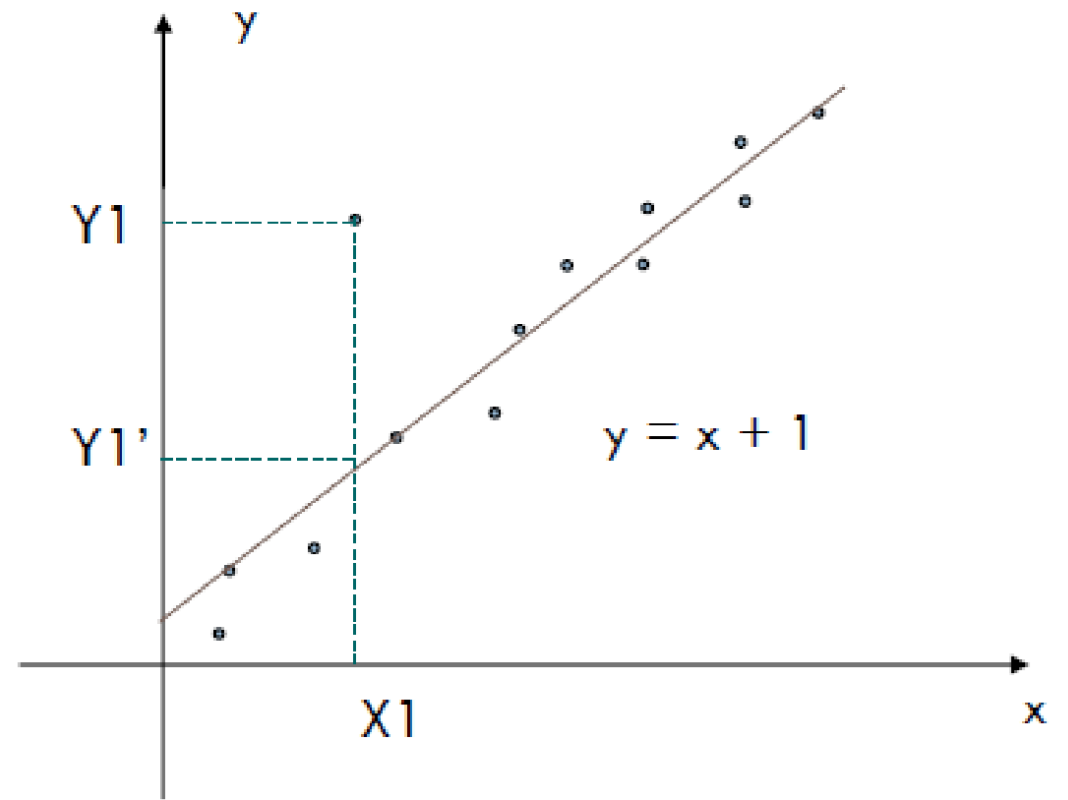
- The regression functions are used to determine the relationship between the dependent variable (target field) and one or more independent variables. The dependent variable is the one whose values you want to predict, whereas the independent variables are the variables that you base your prediction on.
- A RegressionModel defines three types of regression models: **linear**, **polynomial**, and **logistic regression**.
- Linear and stepwise-polynomial regression are designed for numeric dependent variables having a continuous spectrum of values. These models should contain exactly one regression table.
- Logistic regression is designed for categorical dependent variables.

For **linear** and **stepwise regression**, the regression formula is:

*Dependent variable = intercept + $\sum_i (\text{coefficient}_i * \text{independent variable}_i) + \text{error}$*

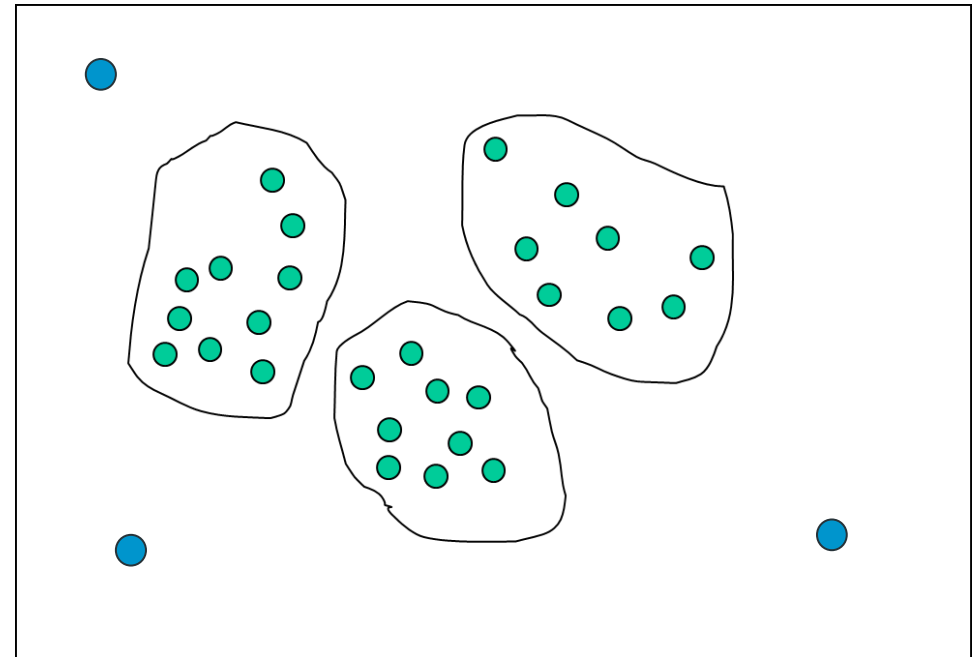
Regression for Data Smoothing

- Replace noisy or missing values by predicted values
- Requires model of attribute dependencies (may be wrong!)
- Can be used for data smoothing or for handling missing data



Clustering for Data Smoothing: Outlier Removal

- Data points inconsistent with the majority of data
- Different outliers
 - Noisy: CEO's salary (-10)
 - In consistent: One's age = 200
- Removal methods
 - Clustering
 - Curve-fitting
 - Hypothesis-testing with a given model



Major Tasks in Data Preprocessing

- Data Cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data Integration
 - Integration of multiple databases, or files
- Data Transformation
 - Normalization and aggregation
- Data Reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data Discretization (for numerical data)

Data Integration

- **Data integration:**
 - Combines data from multiple sources into a coherent store
 - Careful integration can help reduce & avoid redundancies and inconsistencies
 - This helps to improve accuracy & speed of subsequent data mining
 - Heterogeneity & structure of data pose great challenges
 - Issues that need to be addressed:
 1. How to match schema & objects from different sources? (**Entity identification problem**)
 2. Are any attributes correlated?
 3. Tuple duplication
 4. Detection and resolution of data value conflicts

Data Integration

Redundancy & Correlation Analysis:

- Redundant data occur often when integration of multiple databases
 - *Object identification*: The same attribute or object may have different names in different databases
 - *Derivable data*: One attribute may be a “derived” attribute in another table, e.g., annual revenue
- Some redundant attributes may be able to be detected by *correlation analysis*.
 - For nominal data or categorical or qualitative data, we use the χ^2 (chi-square) test.
 - For numeric attributes or quantitative data, we can use the correlation coefficient and covariance, both of which assess how one attribute’s values vary from those of another.
- Careful integration of the data from multiple sources may help reduce/avoid redundancies and inconsistencies and improve mining speed and quality

Why correlation a useful metric?

- Correlation can help in **predicting** one quantity from another
- Correlation can **indicate the presence of a causal relationship (may not be true in all cases)**
- Correlation is used as a **basic quantity** and foundation for many other modeling techniques
- More formally, correlation is a **statistical measure** that describes the association between **random variables**.
- There are several methods for calculating the correlation coefficient, each measuring different types of strength of association

Correlation Analysis (Numeric data)

- Evaluate correlation between 2 attributes, A & B, by computing **correlation coefficient (Pearson's product moment coefficient)**

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

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

$\Sigma(a_i b_i)$ 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)

Correlation Analysis (Numeric data)

- Pearson correlation coefficient is most widely used.
 - Measures the linear association between continuous variables.
 - Ranges between +1 and -1 $-1 \leq r_{A,B} \leq +1$

Correlation Analysis (Numeric data)

- If the resulting value is greater than 0, then
 - A and B are *positively correlated*, meaning that the values of A increase as the values of B increase
 - Higher the value, the stronger the correlation
 - Higher value may indicate that A (or B) may be removed as a redundancy
- If the resulting value is equal to 0, then
 - A and B are *independent* and there is no correlation between them
- If the resulting value is less than 0, then
 - A and B are *negatively correlated*, where the values of one attribute increase as the values of the other attribute decrease

Covariance Analysis (Numeric Data)

- 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

$$\{(a_1, b_1), \dots, (a_n, b_n)\}$$

- Mean values of A and B , respectively, are also known as the **expected values** on A and B , that is

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

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

- **Covariance** between A and B is defined as $Cov(A, B) = E((A - \bar{A})(B - \bar{B})) = \frac{\sum_{i=1}^n (a_i - \bar{A})(b_i - \bar{B})}{n}.$

Covariance Analysis(Numeric Data)

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

- Also, for simplified calculations

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

- **Positive covariance:** If $Cov(A, B) > 0$, then A and B both tend to be larger than **their expected values**
- **Negative covariance:** If $Cov(A, B) < 0$ then if A is larger than its expected value, B is likely to be smaller than its **expected value**
- **Independence:** $Cov(A, B) = 0$ but it may not be true: Some pairs of random variables may have a covariance of 0 but are not independent.

Covariance Analysis (Numeric Data)

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

- Table shows stock prices of two companies at five time points. If the stocks are affected by same industry trends, determine whether their prices rise or fall together?

Covariance Analysis(Numeric Data)

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

and

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

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. ■

Comparison of Correlation and Covariance

Basis for comparison	Covariance	Correlation
Definition	Covariance is an indicator of the extent to which 2 random variables are dependent on each other. A higher number denotes higher dependency.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	The value of covariance lies in the range of $-\infty$ and $+\infty$.	Correlation is limited to values between the range -1 and +1
Change in scale	Affects covariance	Does not affect the correlation
Unit-free measure	No	Yes
Applications	Biology, Astronomy, Prediction of amount investment on different assets in financial markets	Time vs Money analysis by a customer on e-commerce websites, weather forecast, pattern recognition, temperature vs water consumption study, etc

Data Value Conflict Detection and Resolution:

- For the same real-world entity, attribute values from different sources may differ
 - e.g. Prices of rooms in different cities may involve different currencies
- Attributes may also differ on the abstraction level, where an attribute in one system is recorded at, say, a lower abstraction level than the “same” attribute in another
- e.g. total sales in one database may refer to one branch of All_Electronics, while an attribute of the same name in another database may refer to the total sales for All_Electronics stores in a given region.
- For a hotel chain, the price of rooms in different cities may involve not only different currencies but also different services (e.g., free breakfast) and taxes.
- When exchanging information between schools, for example, each school may have its own curriculum and grading scheme.
- To resolve, data values have to be converted into a consistent form

Major Tasks in Data Preprocessing

- Data Cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data Integration
 - Integration of multiple databases, or files
- Data Transformation
 - Normalization and aggregation
- Data Reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data Discretization (for numerical data)

Data Transformation by Normalization

- The measurement unit used can affect the data analysis.
- For example, changing measurement units from meters to inches for height, or from kilograms to pounds for weight, may lead to very different results.
- In general, expressing an attribute in smaller units will lead to a larger range for that attribute, and thus tend to give such an attribute greater effect
- To help avoid dependence on the choice of measurement units, the data should be normalized or standardized.
- This involves transforming the data to fall within a smaller or common range such as $[-1,1]$ or $[0.0, 1.0]$.
- Normalizing the data attempts to give all attributes an equal weight.

Data Transformation by Normalization

- Normalization is particularly useful for classification algorithms involving neural networks or distance measurements such as nearest-neighbor classification and clustering.
- There are many methods for data normalization. We will study:
 - min-max normalization
 - z-score normalization
 - Normalization by decimal scaling
- **Aggregation:** summarization, data cube construction

Data Transformation by min-max Normalization

- Min-max normalization performs a linear transformation on the original data.
- Let A be a numeric attribute with n observed values, v_1, v_2, \dots, v_n .
- 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 $[\text{new_min}_A, \text{new_max}_A]$ by computing

$$v'_i = \frac{v_i - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A. \quad (3.8)$$

- 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 .

Example **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 $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0 - 0) + 0 = 0.716$. ■

Exercise

- Define Normalization.
- What is the value range of min-max. Use min-max normalization to normalize the following group of data: 8,10,15,20.

- Solution:
$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Marks	Marks after Min-Max normalization
8	
10	
15	
20	

Exercise

- Define Normalization.
- What is the value range of min-max. Use min-max normalization to normalize the following group of data: 8,10,15,20.

- Solution:
$$v' = \frac{v - \min_A}{\max_A - \min_A} (\text{new_max}_A - \text{new_min}_A) + \text{new_min}_A$$

Marks	Marks after Min-Max normalization
8	0
10	0.16
15	0.58
20	1

Data Transformation by 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

$$v'_i = \frac{v_i - \bar{A}}{\sigma_A}, \quad (3.9)$$

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

- This method of normalization is useful when the actual minimum and maximum of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Example **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 $\frac{73,600 - 54,000}{16,000} = 1.225$. ■

Normalization by Decimal Scaling

- We move the decimal point of values of the attribute.
- This movement of decimal points totally depends on the maximum value among all values in the attribute
- Normalized Attribute $V' = V_i / 10^j$
- For example (85, 36, 9),
 - Max digits in max value, $j = 2$
 - Normalized Attribute (0.85, 0.36, 0.09)

Major Tasks in Data Preprocessing

- Data Cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data Integration
 - Integration of multiple databases, or files
- Data Transformation
 - Normalization and aggregation
- Data Reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data Discretization (for numerical data)

Data Reduction Strategies

- Data is too big to work with
- Data reduction
 - Obtain a reduced representation of the data set that is much smaller in volume but yet produce the same (or almost the same) analytical results
- Why data reduction? — A database/data warehouse may store terabytes of data. Complex data analysis may take a very long time to run on the complete data set.

Data Reduction Strategies

- Data Cube Aggregation
- Dimensionality Reduction
- Data Compression
- Numerosity Reduction
- Discretization and Concept Hierarchy Generation

Data Cube Aggregation

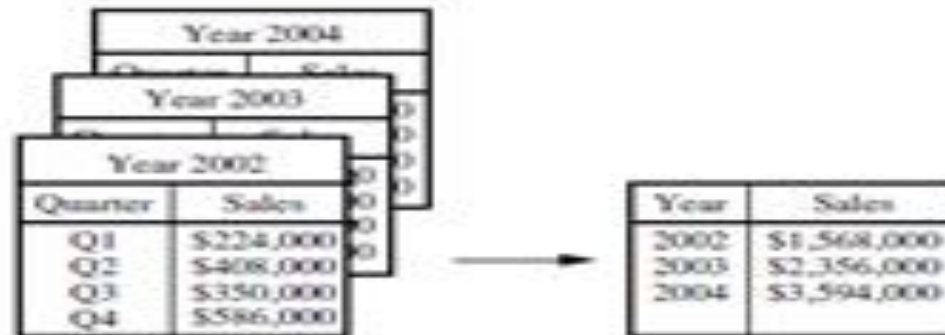


Figure 2.13 Sales data for a given branch of *AllElectronics* for the years 2002 to 2004. On the left, the sales are shown per quarter. On the right, the data are aggregated to provide the annual sales.

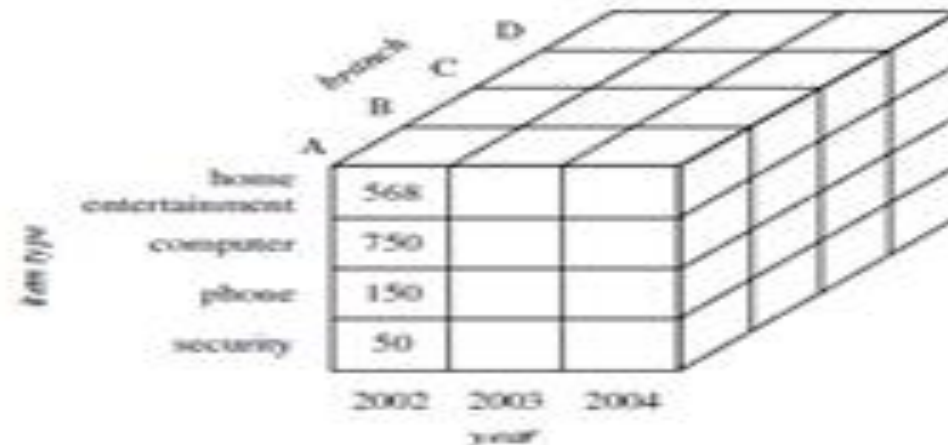


Figure 2.14 A data cube for sales at *AllElectronics*.

Data Cube Aggregation

- Multiple levels of aggregation in data cubes
 - Further reduce the size of data to deal with
- Reference appropriate levels
 - Use the smallest representation capable to solve the task
- Queries regarding aggregated information should be answered using data cube, when possible

Dimensionality Reduction

- A process of reducing the number of random variables or attributes under consideration.
- Data encoding or transformations are applied to obtain a reduced or compressed representation of the original data.
- Include wavelet transforms and principal components analysis (PCA) which transform or project the original data onto a smaller space.
 - Attribute subset selection/Feature subset selection/feature creation: Irrelevant, weakly relevant, or redundant attributes or dimensions are detected and removed.

Attribute Subset Selection

- How can we find a ‘good’ subset of the original attributes?”
- For n attributes, there are 2^n possible subsets.
- An exhaustive search for the optimal subset of attributes can be prohibitively expensive, especially as n and the number of data classes increase.
- Therefore, heuristic methods that explore a reduced search space are commonly used for attribute subset selection.
- The “best” (and “worst”) attributes are typically determined using **tests of statistical significance**, which assume that the attributes are independent of one another.
- Many other attribute evaluation measures can be used such as the **information gain and Gini index** measures used in building decision trees for classification.

Heuristic (Greedy) methods for attribute subset selection

1. Stepwise Forward Selection:

- Starts with an empty set of attributes as the reduced set
- Best of the relevant attributes is determined and added to the reduced set
- In each iteration, best of remaining attributes is added to the set

2. Stepwise Backward Elimination:

- Here all the attributes are considered in the initial set of attributes
- In each iteration, worst attribute remaining in the set is removed

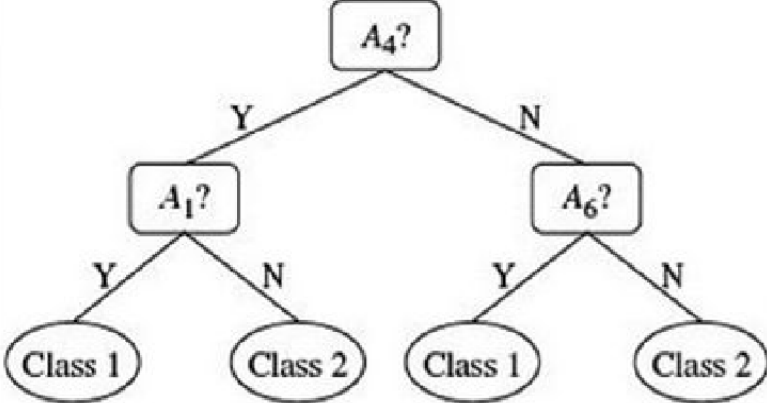
3. Combination of Forward Selection and Backward Elimination:

- Stepwise forward selection and backward elimination are combined
- At each step, the procedure selects the best attribute and removes the worst from among the remaining attributes

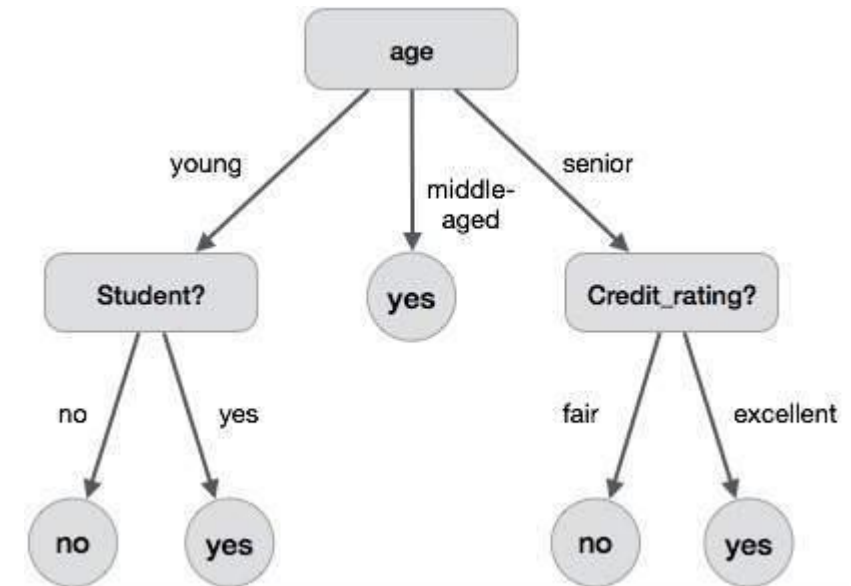
4. Decision Tree Induction:

- This approach uses decision tree for attribute selection.
- It constructs a flow chart like structure having nodes denoting a test on an attribute.
- Each branch corresponds to the outcome of test and leaf nodes is a class prediction.
- The attribute that is not the part of tree is considered irrelevant and hence discarded

Example of Decision Tree Induction

Forward selection	Backward elimination	Decision tree induction
<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p> <p>Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ \Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>	<p>Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$</p>  <p>\Rightarrow Reduced attribute set: $\{A_1, A_4, A_6\}$</p>

Decision Tree: whether a customer at a company is likely to buy a computer or not.



Data Reduction: Numerosity Reduction

- Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data.
- Dimensionality reduction and numerosity reduction techniques can be considered forms of data compression.
- **Parametric methods**
 - Assume the data fits some model, estimate model parameters, store only the parameters, and discard the data (except possible outliers)
- **Non-parametric methods**
 - Do not assume models
 - Major families: histograms, clustering, sampling

Parametric Methods: e.g. Regression and Log-Linear Models

- Linear regression: Data are modeled to fit a straight line: $Y = \alpha + \beta X$
- Multiple regression: allows a response variable y to be modeled as a linear function of multidimensional feature vector (predictor variables)

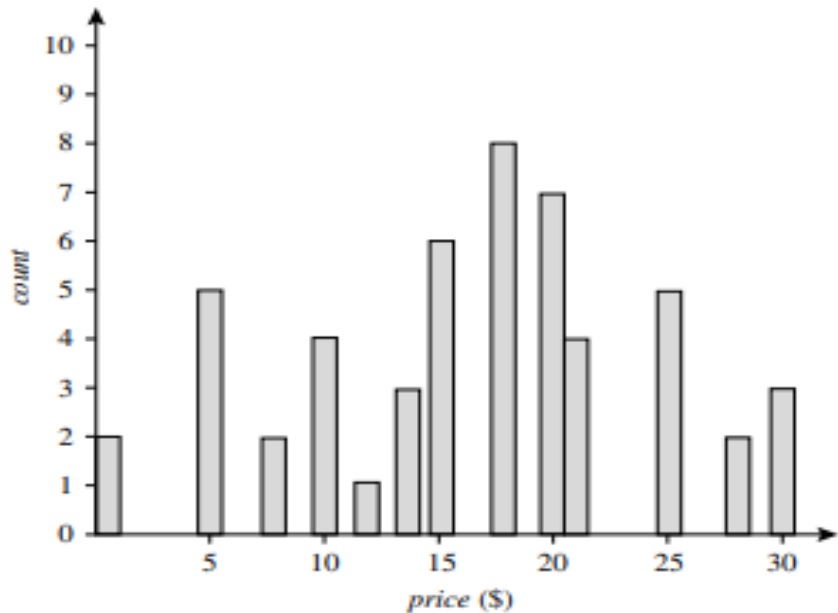
$$Y = b_0 + b_1 X_1 + b_2 X_2.$$

- Log-linear model:

Non-Parametric Methods: 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 represent continuous ranges for the given attribute.
- Example- 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, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

Non-Parametric Methods: Histograms

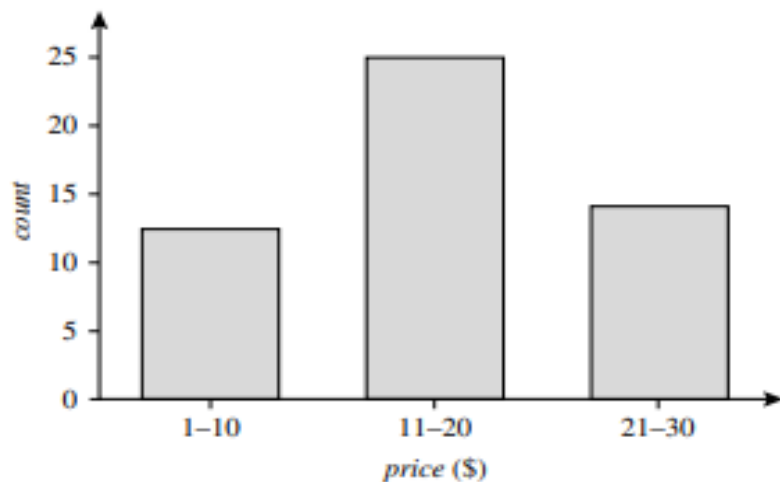


A histogram for *price* using singleton buckets—each bucket represents one price-value/frequency pair.

Example- 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, 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.

Non-Parametric Methods: Histograms

- To further reduce the data, it is common to have each bucket denote a continuous value range for the given attribute.
- In Figure below, each bucket represents a different \$10 range for price.



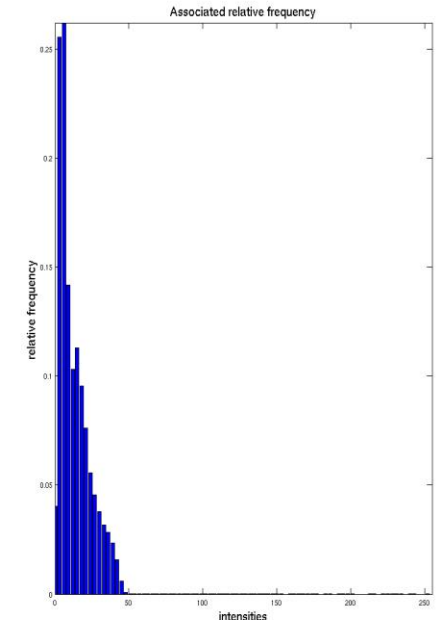
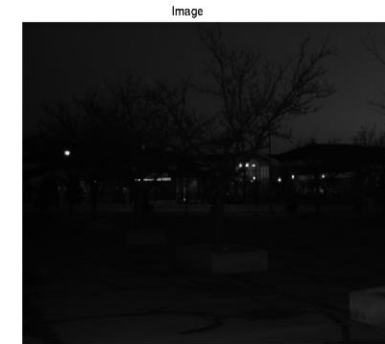
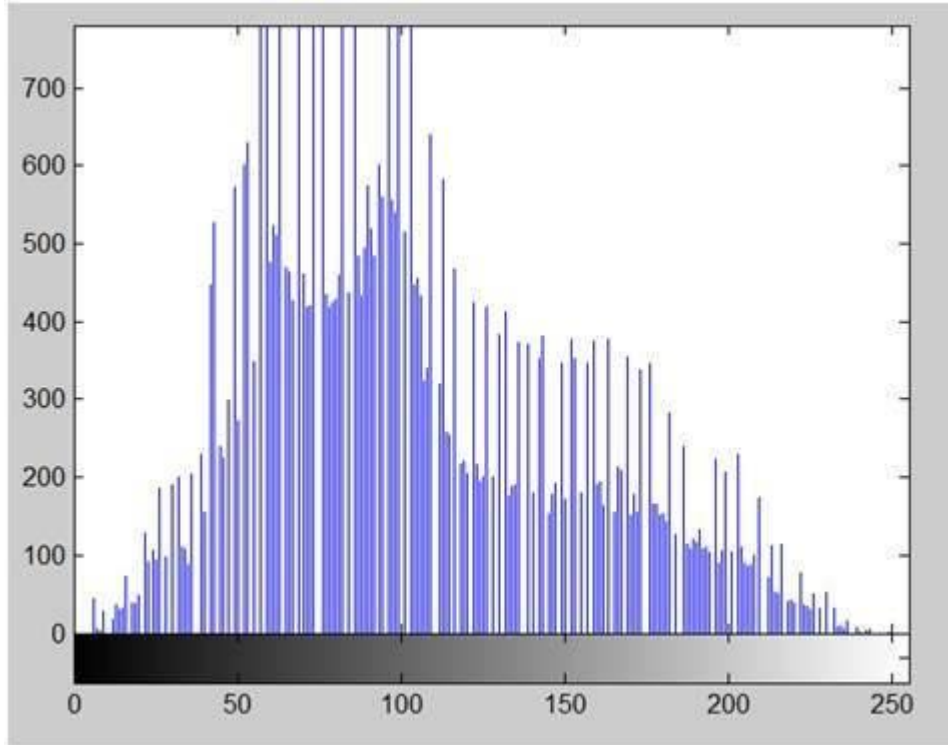
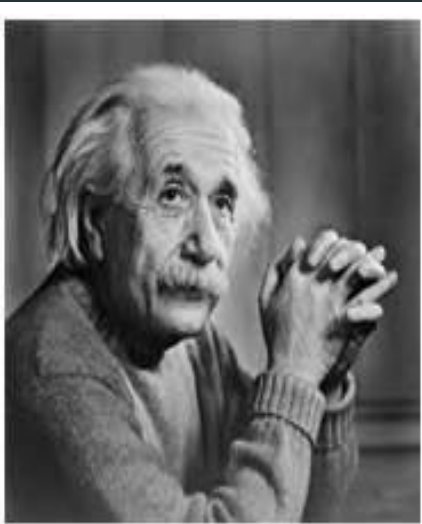
Example- Histograms. The following data are a list of AllElectronics prices for commonly sold items (rounded to the nearest dollar). The numbers have been sorted: 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

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

Non-Parametric Methods: Histograms

- “How are the buckets determined and the attribute values partitioned?” There are several partitioning rules, including the following:
 - Equal-width histogram: The width of each bucket range is uniform (e.g., the width of \$10 for the buckets in Figure).
 - Equal-frequency(or equal-depth) 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).
- Histograms are highly effective at approximating both sparse and dense data, as well as highly skewed and uniform data.
- The histograms can be extended for multiple attributes.

Histogram of an image : Application



- In an image histogram, the **x axis shows the gray level intensities** and the **y axis shows the frequency of these intensities**.

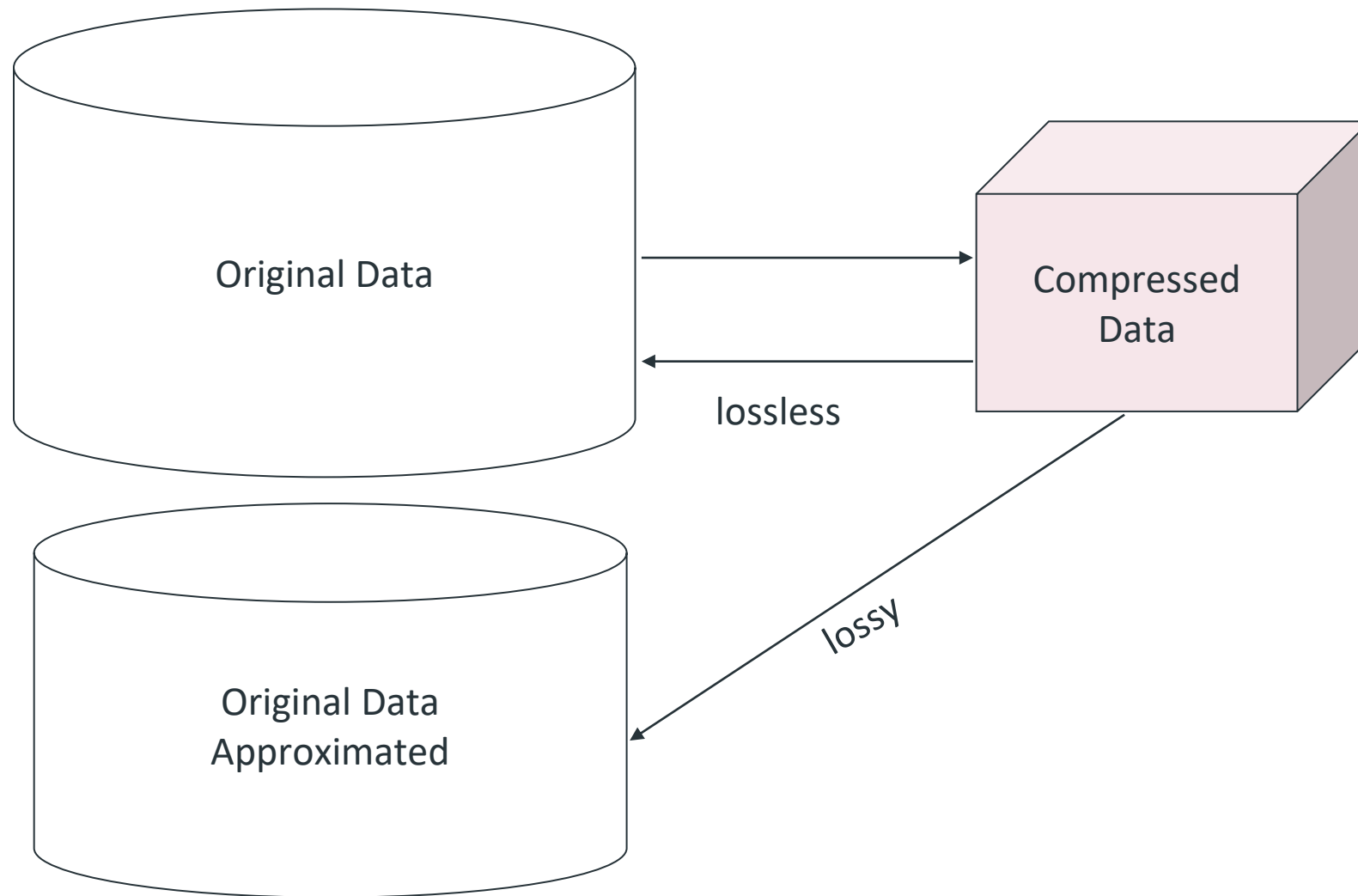
8 bits/per pixel image ---- 256 level of gray or its shades.

As most of the bars that have high frequency lies in the first half portion which is the darker portion. That means that the image we have got is darker

Data Compression

- String compression
 - There are extensive theories and well-tuned algorithms
 - Typically **lossless**
 - But only limited manipulation is possible
- Audio/video, image compression
 - Typically **lossy** compression, with progressive refinement
 - Sometimes small fragments of signal can be reconstructed without reconstructing the whole

Data Compression



Clustering

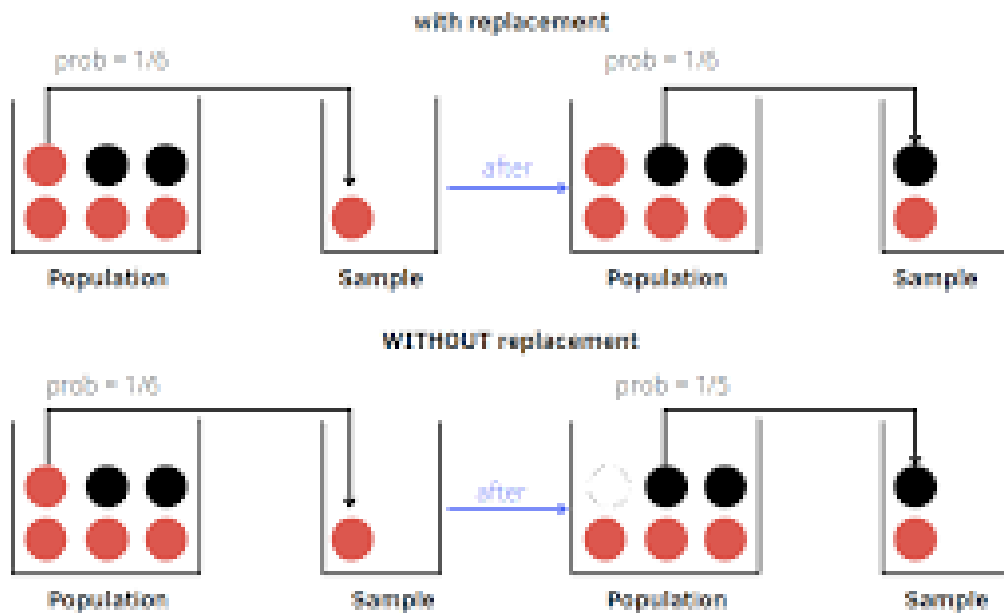
- **Partition** data set into clusters, and store cluster representation only
- **Quality** of clusters measured by their **diameter** (max distance between any two objects in the cluster) or centroid distance (avg. distance of each cluster object from its centroid)
- Can be very effective if the data is not smeared
- Can have hierarchical clustering (possibly stored in multi-dimensional index tree structures)
- There are many choices of clustering definitions and clustering algorithms (further details later)

Sampling

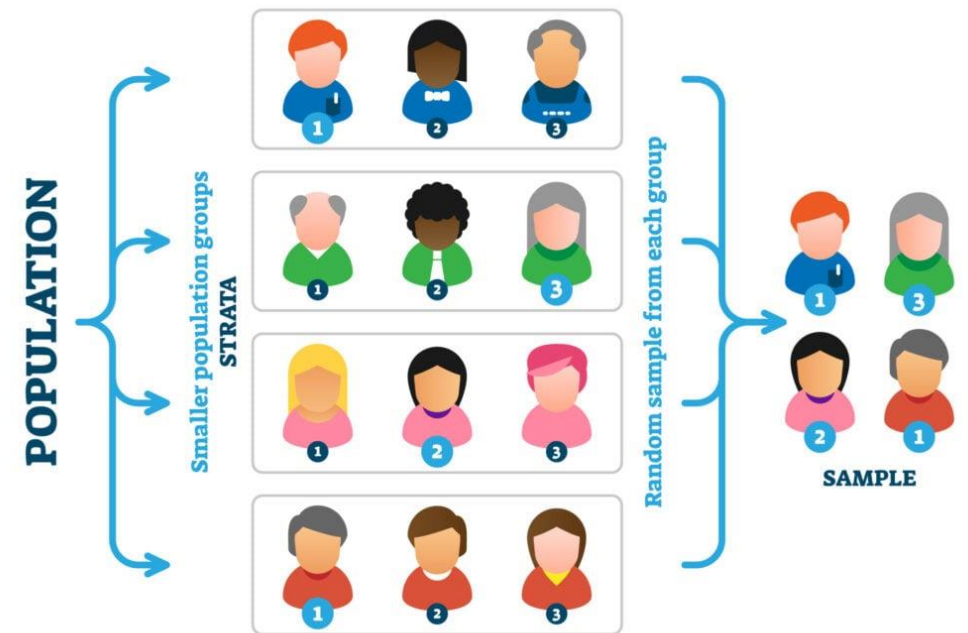
- **Simple random sampling**
 - There is an equal probability of selecting any particular item
 - May have very poor performance in the presence of skew
- **Sampling without replacement**
 - Once an object is selected, it is removed from the population
- **Sampling with replacement**
 - A selected object is not removed from the population
- **Stratified sampling:**
 - Partition the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
 - Used in conjunction with skewed data

Types of Sampling

Random Sampling with/without Replacement

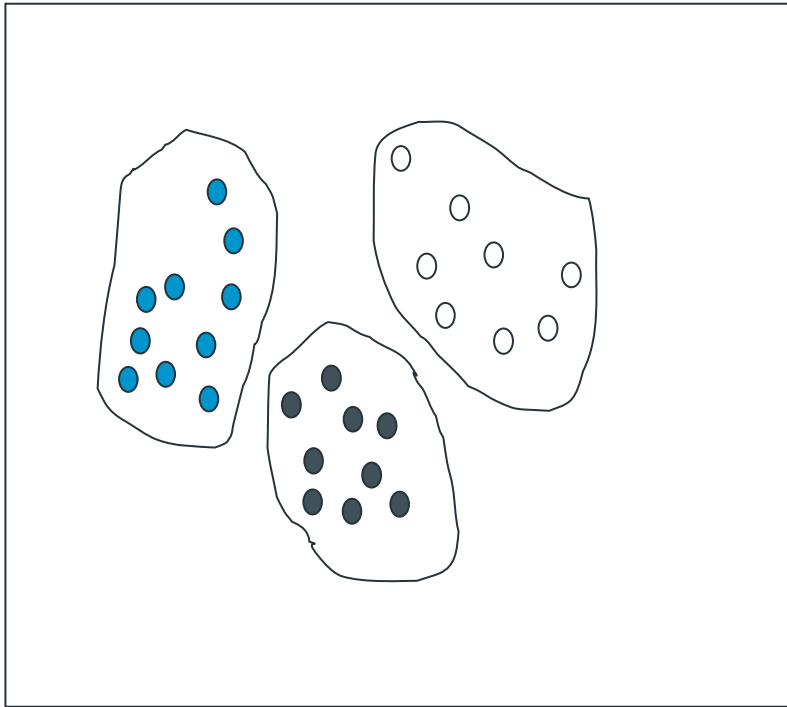


STRATIFIED SAMPLING

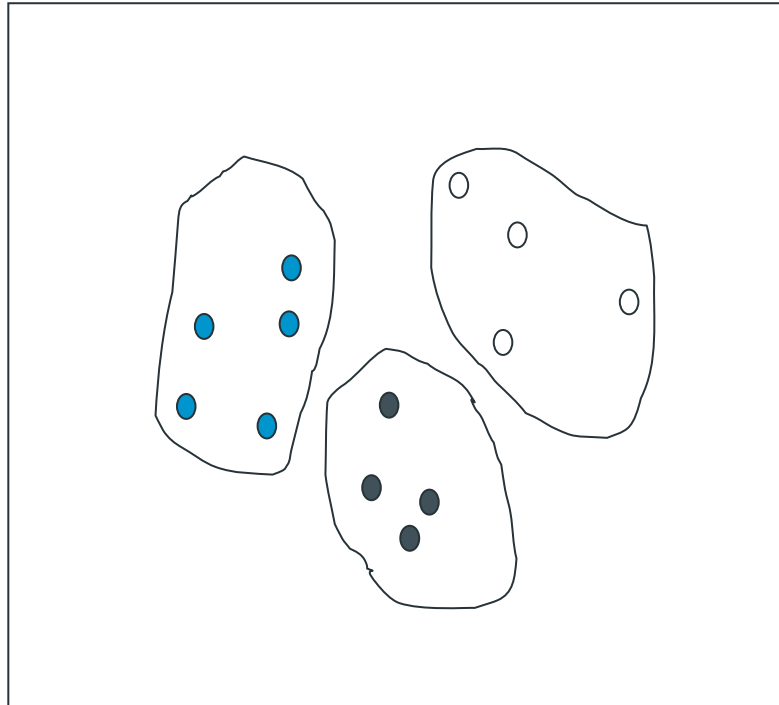


Sampling

Raw Data



Cluster/Stratified Sample



Major Tasks in Data Preprocessing

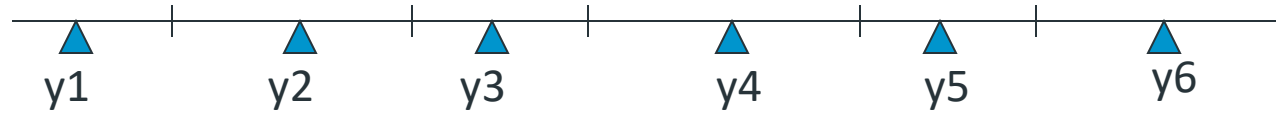
- Data Cleaning
 - Fill in missing values, smooth noisy data, identify or remove outliers and noisy data, and resolve inconsistencies
- Data Integration
 - Integration of multiple databases, or files
- Data Transformation
 - Normalization and aggregation
- Data Reduction
 - Obtains reduced representation in volume but produces the same or similar analytical results
- Data Discretization (for numerical data)

Discretization and Concept Hierarchies

- Discretization
 - Reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.
- Concept Hierarchies
 - Reduce the data by collecting and replacing low level concepts (such as numeric values for the attribute age) by higher level concepts (such as young, middle-aged, or senior).

Discretization

- Three types of attributes:
 - Nominal — values from an unordered set
 - Ordinal — values from an ordered set
 - Continuous — real numbers
- Discretization/Quantization:
 - divide the range of a continuous attribute into intervals



- Some classification algorithms only accept categorical attributes.
- Reduce data size by discretization
- Prepare for further analysis

Discretization and Concept Hierarchies : Numerical data

- Hierarchical and recursive decomposition using:
 - Binning (data smoothing)
 - Histogram analysis (numerosity reduction)
 - Clustering analysis (numerosity reduction)
 - Entropy-based discretization
 - Segmentation by natural partitioning

Summary

- **Data quality:** accuracy, completeness, consistency, timeliness, believability, interpretability
- **Data cleaning:** e.g. missing/noisy values, outliers
- **Data integration** from multiple sources:
 - Entity identification problem
 - Remove redundancies
 - Detect inconsistencies
- **Data reduction**
 - Cube Aggregation
 - Dimensionality reduction
 - Numerosity reduction
 - Data compression
- **Data transformation and data discretization**
 - Normalization
 - Concept hierarchy generation

Reference

- **Data Mining: Concepts and Techniques**, Jiawei Han, Micheline Kamber, and Jian Pei, 3rd edition