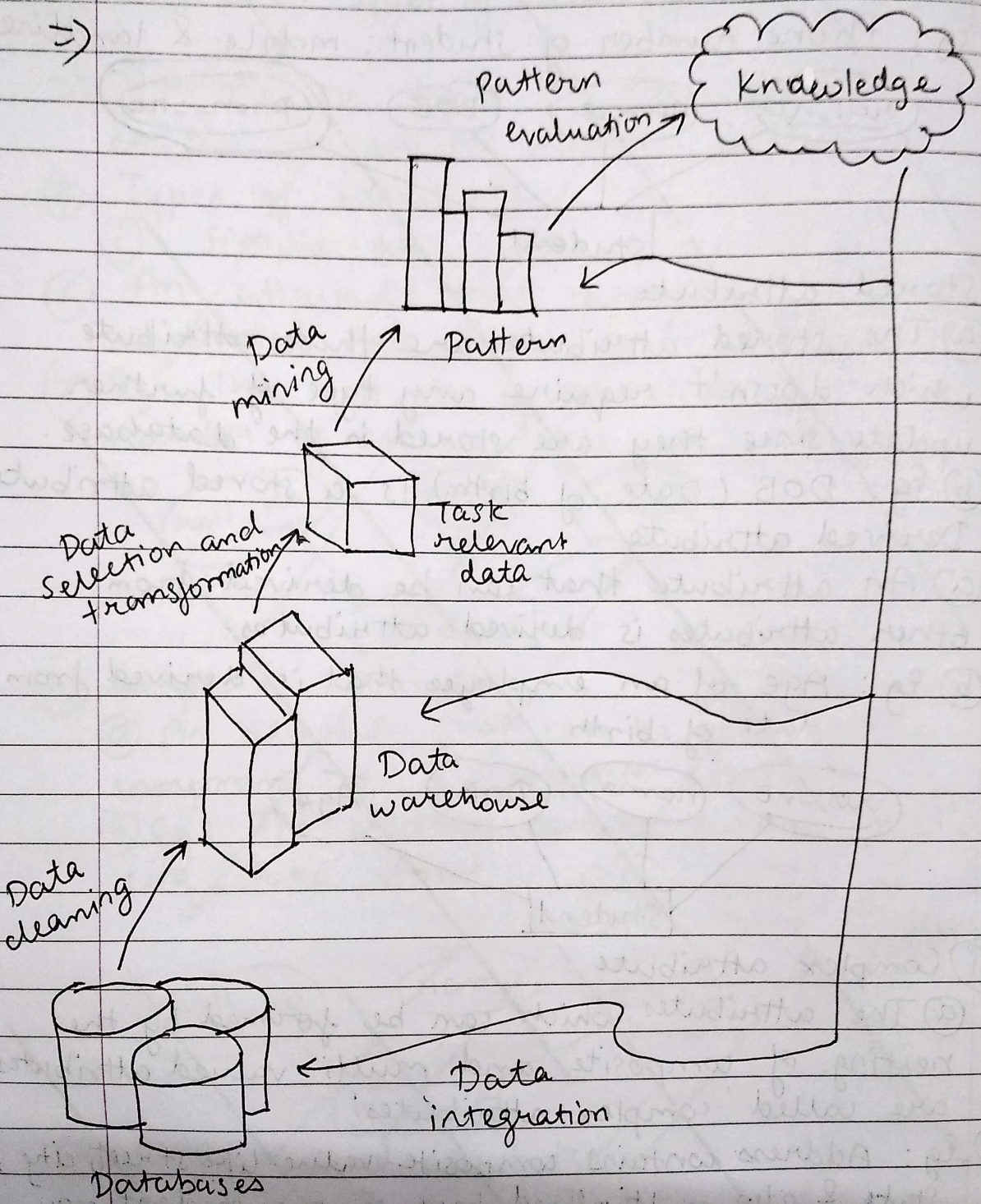
Module 2

(Q1) KDD process (knowledge discovery in database)

=>



(1) The main objective of the KDD process is to extract information from data in the context of large databases.

(2) It helps organizations to analyze data effectively for better decision making.

(3) Steps in KDD process

(i) Data cleaning

(a) In this step, noisy, irrelevant or missing data is removed.

(b) Techniques include handling missing values, removing errors & using transformation tools for consistency.

(ii) Data integration

(a) The data from multiple sources is combined in a common source.

(b) Data integration is done using tools like data migration tools, ETL process.

(iii) Data selection

(a) The process where data relevant to the analysis is decided & retrieved from the data collection.

(iv) Data transformation

(a) The process of transforming data into suitable format required for the mining

(v) Pattern evaluation

(a) Pattern evaluation identifies strictly increasing patterns which represent knowledge based on given measures.

(vi) Data mining

(a) The techniques that are applied to extract potentially useful patterns

(b) Techniques include classification, clustering or regression.

(vii) Knowledge representation

(a) It involves presenting the results in the form of graphs or reports.

Page No.
Date

Q2) Types of attributes

=> (1) Attributes are features or variables that describe the data.
(2) They are described into two main types:
(i) Qualitative attributes
(ii) Quantitative attributes
(3) Qualitative attributes
(a) Qualitative attributes describe non-numerical data.
(b) Nominal data (N):
(i) Nominal attributes are those that represent names or categories.
(ii) Eg: A person's eye color is a nominal attribute because there is no ranking of eye color (blue is not 'better' than brown)
(c) Binary attributes (B):
(i) These are the attributes which can only take on two values i.e. 0 and 1, or, true or false, Yes or No, etc.
(ii) Eg: The attribute 'cancer detection' can have values Yes (presence of cancer) and No. (absence of cancer)
(d) Ordinal attributes (O)
(i) Ordinal attributes are attributes that have a specific order or ranking.
(ii) These attribute indicate a rank or position but do not quantify the difference between ranks.
(iii) Eg: The attribute 'grade' can have values such as A, B, C, D, E and F, representing the ranking of performance.

(4) Quantitative attributes

(a) Quantitative attributes describe numerical data and are used to measure quantities

(b) Discrete attributes

(i) Discrete attributes have a finite or countable number of values.

(ii) These values can be numeric or categorical but are always different & seperate.

(iii) Eg: The attribute 'zip-code' can have values like 400091 or 111400

(c) Continuous attributes

(i) Continuous attributes have an infinite number of possible values within a given range.

(ii) Eg: The attribute 'height' can have values like 5.2, 5.6, 6.0, 6.4, etc.

(d) Interval attributes

(i) Interval-scaled attributes are numerical attributes where the differences between values are meaningful, which do not have a true zero point

(ii) These attributes are continuous measurements on linear scale.

(iii) Eg: Temperature measured in celsius (°C)

(e) Ratio attributes

(i) Ratio-scaled attributes are numerical attributes that have a true zero point.

(ii) Eg: The attribute 'Age' can have values like 20 years and 40 years. Here, we can say that a person who is 40 years old is twice as old as someone who is 20 years old.

Q3) Data Pre-processing

=> (1) Data pre-processing is a critical step in the data mining process
(2) It involves cleaning, transforming, integrating and preparing raw data to make it suitable for analysis.
(3) Steps of data pre-processing:
(i) Data cleaning
(a) Data cleaning focuses on correcting errors and removing inconsistencies to ensure the dataset is accurate and complete.
(b) Data cleaning involves
(1) Handling missing values
(2) Handling noisy data
(ii) Data integration
(a) The data from multiple sources is combined into a single source
(b) Data integration is done using tools like data migration tools, ETL process.
(c) Common methods include:
(1) Record linkage
(2) Data fusion
(iii) Data Transformation
(a) The process of transforming data into suitable format which is required for mining
(b) Data transformation includes:
(1) Normalization
(2) Discretization
(3) Concept hierarchy generation
(iv) Data reduction
(a) Data reduction reduces the dataset's size

while preserving essential information, making analysis more efficient.

(b) Data reduction includes : Feature selection, Feature extraction, Sampling, Compression.

(v) Data Discretization

(a) Discretization converts continuous data into discrete categories to simplify analysis.

(b) Common techniques used in data discretization is clustering.

(vi) Data normalization.

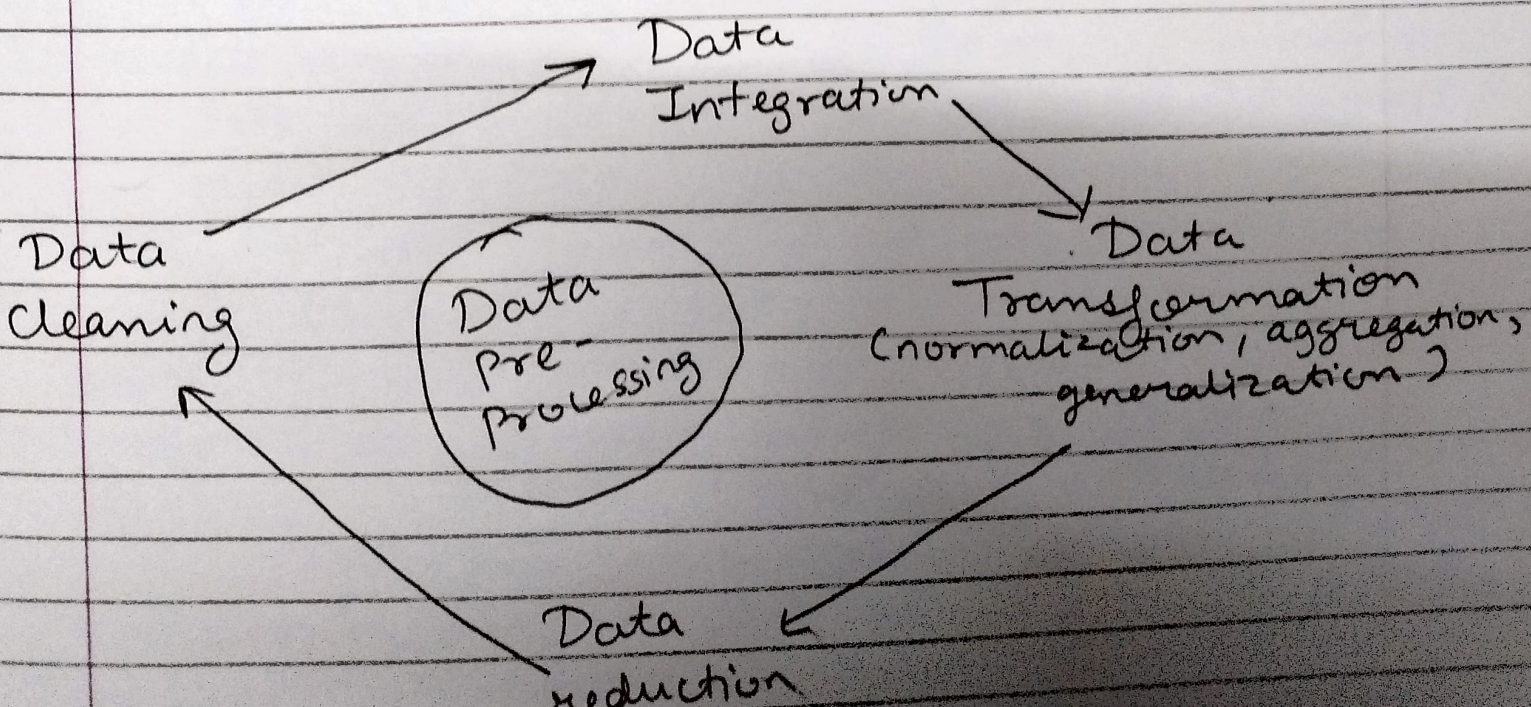
(a) Normalization scales data to eliminate differences in units or magnitude, making it comparable.

(b) Common techniques include :

(1) Min - max normalization

(2) Z - score normalization

(3) Decimal scaling



Q4. Applications of Data Mining.

=>

Data mining is the process of discovering meaningful patterns, trends, and relationships within large datasets. It is widely used across various industries and domains to extract valuable insights and support decision-making.

1. Healthcare and Medicine:

- **Usage:** Data mining is applied to analyze patient records, predict disease outbreaks, and personalize treatment plans.
- **Example:** Predicting the likelihood of heart disease in patients using historical medical data.

2. Retail and E-commerce:

- **Usage:** Retailers use data mining for market basket analysis, customer segmentation, and inventory management.
- **Example:** Amazon's recommendation system, which suggests products based on past purchases and browsing history.

3. Banking and Finance:

- **Usage:** Banks and financial institutions apply data mining to detect fraud, assess credit risk, and optimize investment strategies.
- **Example:** Identifying fraudulent transactions by detecting unusual spending patterns.

4. Education:

- **Usage:** Educational institutions use data mining to track student performance, optimize curriculums, and predict dropout rates.
- **Example:** Predicting students at risk of failing based on attendance and test scores.

• 5. Sports and Entertainment:

- **Usage:** Data mining is used to analyze player performance, predict match outcomes, and enhance fan engagement.
- **Example:** Using player statistics to draft the best team in fantasy leagues or professional tournaments.

6. Manufacturing:

- **Usage:** Data mining is used to improve production efficiency, predict equipment failures, and manage supply chains.
- **Example:** Predicting when a machine might fail to enable proactive maintenance and reduce downtime.

7. Marketing and Customer Relationship Management (CRM):

- **Usage:** Marketers use data mining to analyze consumer behavior, target advertisements, and measure campaign effectiveness.
- **Example:** Social media platforms like Facebook using data mining to deliver targeted ads based on user preferences.

8. Transportation and Logistics:

- **Usage:** Data mining optimizes routes, predicts traffic patterns, and improves delivery times in logistics.
- **Example:** UPS using data mining to optimize delivery routes and reduce fuel consumption.

9. Government and Public Services:

- **Usage:** Governments use data mining for tax fraud detection, crime prevention, and disaster management.
- **Example:** Analyzing crime data to predict high-risk areas and allocate resources effectively.

10. Social Media Analysis:

- **Usage:** Social media platforms mine user data to understand trends, sentiments, and user engagement.
- **Example:** Twitter using sentiment analysis to gauge public opinion on events or products.

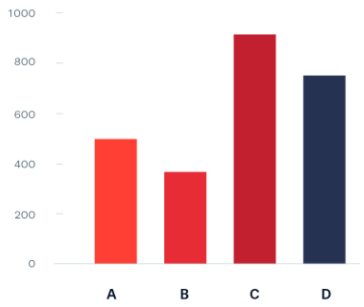
Q5. Explain different data visualization techniques

=>

Data visualization refers to representing data graphically or visually to identify patterns, trends, and insights more effectively. Various techniques are used depending on the type and complexity of the data, and the purpose of the visualization.

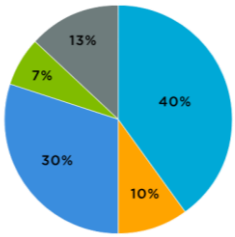
1. Bar Chart

- **Purpose:** Compare categories or groups.
- **Description:** A bar chart represents data using rectangular bars where the length or height of the bar is proportional to the value. It can be horizontal or vertical.
- **Example:** Comparing sales of different products in a quarter.



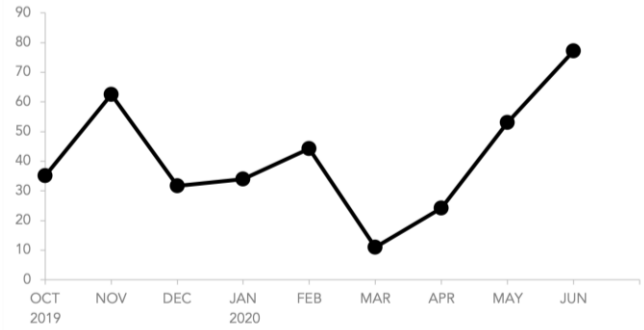
2. Pie Chart

- **Purpose:** Show proportions or percentages.
- **Description:** A circular chart divided into slices, where each slice represents a proportion of the whole.
- **Example:** Showing the percentage of revenue contribution from different regions.



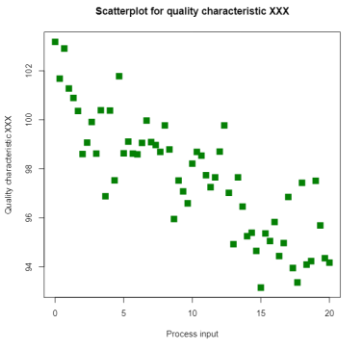
3. Line Chart

- **Purpose:** Visualize trends over time.
- **Description:** A line chart uses points connected by lines to represent data changes over continuous intervals.
- **Example:** Tracking the stock price of a company over a year.



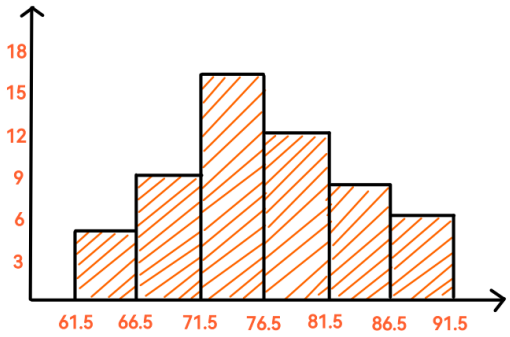
4. Scatter Plot

- **Purpose:** Show relationships between two variables.
- **Description:** A scatter plot uses dots to represent individual data points on an X and Y axis, revealing correlations or trends.
- **Example:** Analyzing the relationship between marketing spend and revenue.



5. Histogram

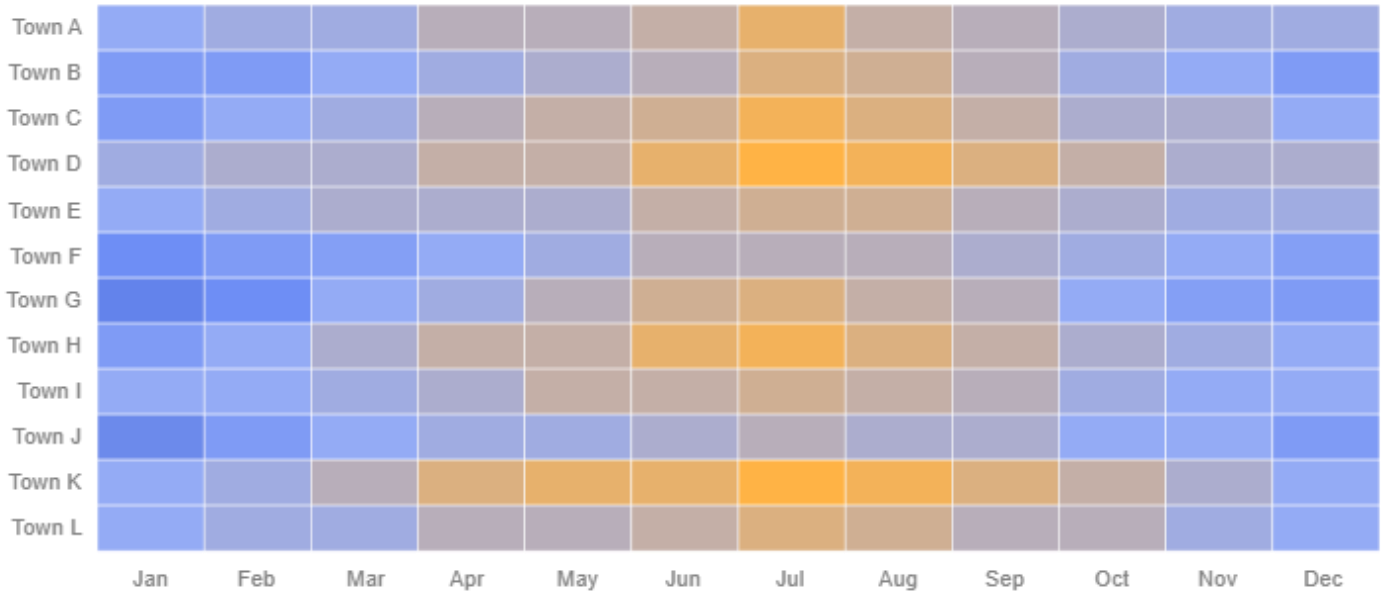
- **Purpose:** Show the frequency distribution of a dataset.
- **Description:** A histogram groups continuous data into bins and represents the frequency of each bin as bars.
- **Example:** Distribution of exam scores among students.



6. Heatmap

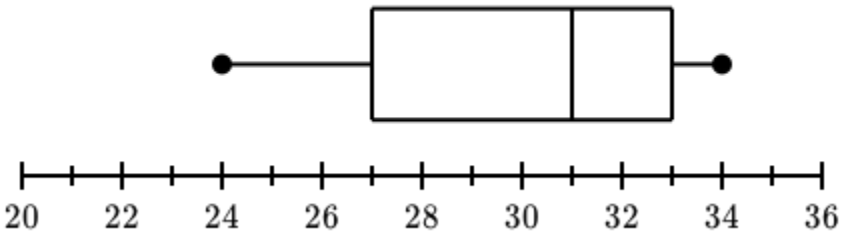
- **Purpose:** Represent data intensity or relationships in a tabular form.

- **Description:** A heatmap uses colors to represent values in a matrix or table. Darker or lighter colors indicate the intensity of data.
- **Example:** Visualizing correlation between multiple variables in a dataset.



7. Box Plot (or Box-and-Whisker Plot)

- **Purpose:** Summarize the distribution of a dataset.
- **Description:** A box plot shows the median, quartiles, and outliers in a dataset. It is useful for identifying variability and spread.
- **Example:** Comparing the distribution of salaries across departments.



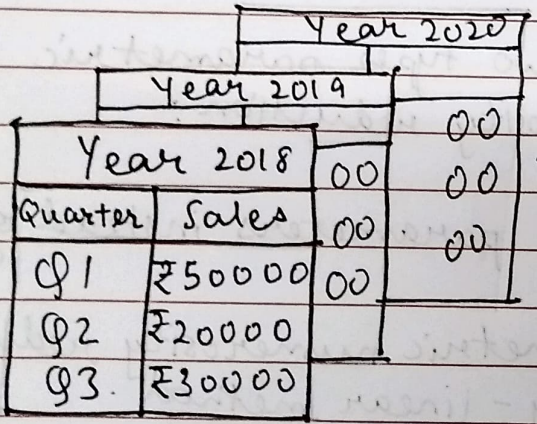
(Q17) Explain the different techniques of data reduction.

=> (a) Data Cube Aggregation

(1) This technique is used to aggregate data in a simpler form.

(2) It is a multi-dimensional aggregation that uses aggregation at various levels of a data cube to represent the original data set, thus achieving data reduction.

(3) Eg: Suppose you have the data of all electronics sales per quarter for the year 2018 to year 2022. If you want to get annual sale per year, you just have to aggregate the data.



Year	Sales
2018	1,00,000
2019	2,50,000
2020	3,90,000

(b) Dimensionality reduction

(1) Dimensionality reduction eliminates the attributes from the data set under consideration, thereby reducing the original volume of data.

(2) It reduces data size as it eliminates outdated or redundant features.

(3) The three methods of dimensionality reduction are:
- (i) Wavelet transform
- (ii) Principal component analysis.
- (iii) Attribute subset selection.

(c) Data compression

(1) Data compression is the process of reducing the number of bits needed to either store or transmit the data.

(2) This data can be text, graphics, video, audio, etc.

(3) It is divided into two types:

(i) Lossless compression: In this the data can be restored from its compressed form.

(ii) Lossy compression: It is not possible to restore the original form from the compressed form.

(d) Numerosity reduction.

(1) The numerosity reduction reduces the original data data volume and represents it in a much smaller form.

(2) This technique includes two types parametric and non-parametric numerosity reduction.

(3) Parametric

(i) It stores only data parameters instead of the original data.

(ii) One method of parametric numerosity reduction is the regression and log-linear method.

(4) Non parametric

(i) A non parametric numerosity reduction technique does not assume any model.

(ii) The non-parametric technique results in a more uniform reduction, irrespective of data size.

(iii) Non-parametric data reduction techniques:

(a) Histogram

(b) Clustering

(c) Sampling