Unit - I

Data management

- Design Data Architecture and manage the data for analysis.
- understand various sources of Data like sensors /signals /GPS etc.
- Data management
- Data Quality (noise, outliers, missing values, duplicate data)
- Data preprocessing & processing.

Data management : Design Data Architecutre and manage the Data for Analysis

Data architecture design is like a delailed planned for how to handle data in a company, showing the steps for gathering storing, accessing and using data, This plan helps keep data neat and well organized. Data management adds to this by taking care of data from start to finish, including collecting, storing, arranging, maintaining and protecting it.

Together, these methods help a bus'ness by making sure data is correct easy to get and dependable, which helps ~~which~~ with. making decisions and running the bosiness smoothly.

Data Architecture Design

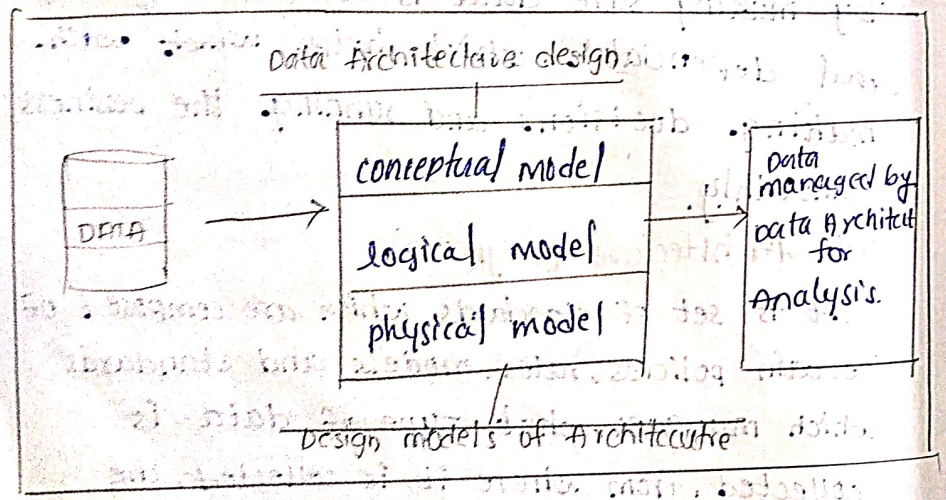
It is set of standards which are composed of certain policies, rules, models and standards which manages, what type of data is collected, from where it is collected, the arrangement of collected data, storing that data, utilizing and securing the data into the systems and data warehouses for

(or)

It is a strategic plan or blueprint that defines how data is collected, organized, stored accessed, integrated, and secured within an organization. It serves as the foundation for all data-related process.

Data architecture also describes the type of data structures applied to manage data and it provides an easy way for data preprocessing.

The data architecture is formed by dividing into 3 essential models and then are combined:



1. conceptual model
- High level representation
- Uses Entity - Relationship (ER) models to show how entities (like customer, Orders) relate.
- Helps stakeholders (non-technical people) understand the system.
- Does not include technical details like data types or storage.

It serves as a blueprint that outlines major entities and their connections.

ex:- In a hospital, the conceptual model shows relationships between patients, doctors and appointments.

2. logical model
- Adds structure to the conceptual model
- shows how data is organized logically (tabels, XML, object classes)
- Developers use it to build databases.
- Contains tables, fields, primary keys, foreign keys. but still independent of actual software/hardware.

The logical model provides a blueprint of how data is organized and how the databases should operate, translating complex systems designs into a readable form for technical developer.

ex:- patient table with columns like ID, Name, Age

3. physical model
- Actual implementation using specific technologies (SQL, Oracle etc)
- Deals with indexes, partitions, storage paths and file formats.
- Used by DBA's (Database Administrators).

Ex:- patient table stored in MySQL with indexing on patient ID

This model is crucial for database administrators who are responsible for the technical deployment and maintenance of the database systems.

- A data architect is responsible for all design, creation, management, deployment of data architecture and defines how data is to be stored and retrieved. Other decisions are made by internal bodies.

Data Management

It is the end-to-end process of collecting, storing maintaining, securing and using data efficiently for analysis and business operations.

Key :- - extracting data from sources
- storing data securely
- ensuring data accessibility, accuracy & consistency
- converting data into usable formats.
- maintaining security and reducing costs.

Responsibilities in Data management

- Accessibility & integrity :-
  ensure authorized users can access accurate, consistent data.
- security
  protect data using encryption, backups access control.
- cost efficiency
  use optimized tools to reduce storage/processing costs.
- support Business Decisions :-
  provide reliable data for planning and strategy.
- Handle Big Data :-
  use tools like Hadoop, Scale, AWS, Tableau for massive datasets.

Relationship : Data Architecture vs Data Management

Data Architecture	Data Management
- framework for data flow	execution of that framework.
- focuses on structure and polices	focuses on operations and security.
- Built by data architects	Managed by IT/data team
- Deals with how data should flow	Deals with how data does flow.

⇒ Both work together to ensure high-quality accessible, and secure data systems.

Benifits of both integrating

- Informed decision making
- Reduced Redundancy
- Automation
- Compliance
- Technology

Real world Applications:-

- Banking sector
- Architecture: Designs secure flow of transaction data
  management: Ensures real time processing and backup of coustmer transactions.
- Health care
  Architecture:- connects patient records across departments
  management: ensures up-to-data records and compliance with laws like HIPAA.

Understand various sources of Data like sensors/signals Gps etc.

In the digital era, data can come from various sources, and these can be:

- Structured (like databases)
- semi-structured (like XML, JSON)
- Unstructured (like audio, video, images)

Understanding the nature and behaviour of these sources is essential for building efficient systems for data collection, storage and analysis.

1. Sensor Data

Sensors are physical devices that detects and measure real-world conditions then convert them into digital data.

example of sensor data:

sensor type	data captured	Applications
Temperature sensor	Temperature (°C/°F)	weather monitoring
motion sensor	Movement status	smart homes, security systems.
Heart Rate sensor	Beats per minute (BPM)	Health wearables ECG monitoring.

characteristics:-

- Real time: provides live updates
- continuous: produces a stream of data.
- Time-series: each reading is linked with a timestamp.

Use case:-

A smart home uses motion sensors to detect movement and turn on lights automatically.

2. Signal Data

Signal data refers to information derived from analog or digital signals, such as sound, electricity or electromagnetic waves.

examples:-

signal type	Data captured	Applications
Audio signals	sound wave amplitude	voice assistants, music processing.
ECG(electrodiogram)	Hearts electrical activity	Heart disease detection.
EEG(electroenphalogram)	Brain wave patterns	sleep research, Neurology.
Radar / Ultrasound	Reflected wave pattern	Navigation.

characteristics:-

- High frequency: captured many times per second.
- Time dependent: values vary over time
- Requires processing: often transformed using techniques like:-
  - FFT (Fast fourier Transform)
  - STFT (short-time fourier Transform)

use case:- EEG signals help doctors analyze brain disorders like epilepsy.

8. Gps Data (Global positioning system)

Gps is a satellite based system that gives data about location, speed and movement.

ex:-

Data type	Description	Application
lattitude & longitude	coordinates of a place	Google maps, Ride apps.
speed & direction	moment behaviour	fleet/vechicle tracking.
Timestamps	when data was recorded	Time-aware analysis (Routes delivary).

Characteristics:-

- spatio-temporal: combines location + time
- periodic or event driven: Data can be sent on schedule or during events.

use case:- Gps in delivary drones tracks real-time routes and drop points.

9. Other sources of Data

source	Description	example.
social media	user-generated text/image/videos.	Tweets/Insta posts.
Logs & Clickstreams	Tracks user activity on apps/web	Login logs, Click paths.
databases	structured internal data	HR records, CRM systems.
APIs	external sources providing data	weather, stock prices.
Drones	multisensor data (videos, audio, Gps)	Surveillence Delivary.

Data Quality (NOISE, OUTLIERS, MISSING values, DUPLICATE DATA)

It refers to the condition or state of data based on various attributes that determine its reliability, accuracy, consistency, and suitability for a specific purpose.

High quality data is accurate, complete and timely ensuring it is free from errors, inconsistencies and bias.

Good data quality supports effective decision-making, analysis and modeling in various fields like business, research and technology.

Key attributes of High-Quality Data.

Attribute	Meaning
Accuracy	– Data reflects the true value of what it represents.
completeness	– No important information is missing
consistency	– No contradictions between datasets.
Timeliness	– Data is up to date when needed.
Reliability	– can be trusted to make business or technical decisions.

1. NOISE in Data

Noise is irrelevent, distorted or corrupt data that masks the true pattern.

Noise refers to meaningless or corrupt data. It often includes typos, special characters, or unwanted symbols.

sources:-

- Typing mistakes (e.g "A@3") – Data transmission errors
- mal functioning sensors.

Ex:- A@3&, Anita, R@hul.

→ problem: "A@3&", "R@hul". contain special characters

→ soln :- Remove noise using cleaning fn (e.g regex, filters)

Characteristics:-

- Reduce data readability
- Affects text analytics or machine learning models.
- often removed using data cleaning techniques.

2. Missing values :-

A missing values is when no data is recorded for a particular field or variable.

Ex:-

Name	Age	score
Anita	20	85
Ravi		92
meena	19	

→ Age missing for Ravi

→ score missing for meena.

Types of missingness :-

Type	explanation
1. MCAR (missing completely at Random)	missing data is not related to any variable e.g:- random nulls Impact:- least bias.
2. MAR (missing at random)	missing depends on other observed variables e.g:- Income missing depending on age Impact - moderate bias.
3. MNAR (Missing Not at Random)	missing depends on unseen factors e.g:- Refusing to report income due to privacy Impact:- Most bias.

Handling Techniques :-

- Fill with mean/median/mode
- Drop rows with too many nulls.
- Use predictive imputation (e.g KNN imputer)

3 Outlier

These are data points that are significantly different from others in the dataset. They can distort statistical analyzes and machine learning models.

Outliers are extreme values that are far away from other data points.

ex:-

student	score
Anita	85
Ravi	1000
Meena	78

→ Ravi score is abnormally high - an outlier.

4. Duplicate Data.

Repeated rows or records in a dataset.

ex:-

Name	score
Ravi	92
Ravi	92
Anita	85.

→ Two rows for Ravi are identical → duplicate.

Why Remove Duplicates!

- Saves storage space.
- Improves model accuracy
- ensures unbiased analysis.

DATA MANAGEMENT

Data management is the systematic process of collecting, storing, organizing, maintaining securing and preparing data for effective analysis and decision making.

key benifits:-

- ensures data quality (accurate, clean, up-to-date)
- enables quick and accurate analysis.
- Helps meet data regulations like GDPR.
- Improves data access for users and systems.

components of Data management

1. Data collection
   - source: sensors, APIs, forms, databases.
   - Goal: capture raw data in real-time or scheduled intervals.
   - ex:- A health app collects heart rate data from smartwatches.
2. Data storage
   - ~~source~~ structured storage formats: SQL, NOSQL, csv
   - storage options:-
     - cloud: AWS, Azure, Google cloud
     - local: Oracle, MySQL, Excel.
3. Data Organization:-
   - Organizing data into tables, schemas. data warehouse / lakes
   - Helps improve retrieval speed & clarity.

ex:- separate tables for students, marks, attendence.

4. Data Cleaning

- Removing noise, duplicates, fixing missing values.
- Ensures data consistency and trustworthiness.

Ex:- Removing invalid timestamps in log data.

5. Data integration:

- combining data from multiple sources
- Resolves data soils and ensures completeness.

Ex:- Merging data from a CRM and an online shopping site.

6. Data Governance.

- Defines policies and standards:
  - who can access data
  - Audit logs
  - Data ownership & ethics.

Ex:- Enforcing privacy rules like encryption or masked data views.

7. Data preparation for Analysis

- Transforming data for easy visualization and ML.

Ex:- Calculating average sales per month.

DATA PREPROCESSING

It is the systematic process of collecting raw data and transforming it into usable meaningful information. This information can be then be analyzed, visualized, or used for decision-making.

- It's usually performed by data scientists or data teams using using software tools and machine learning models.

Why is Important:-

- converts unorganized raw data into structured insights
- make data understandable & actionable.
- Helps in automated decision-making & real-time response.
- Supports data storage, compliance and future use.

6 steps in Data preprocessing

1. Data collection
   - Gather raw data from data lakes, warehouses, sensors, apps, etc.
   - The quality of data depends on how reliable source is.
2. Data preprocessing.
   - Raw data is cleaned and organized
   - Remove noise - fix missing values - Eliminate duplica
3. Data Input
   - Clean data is entered into tools or

4. Data preprocessing
- Algorithms or scripts process the data:
  - machine learning for pattern recognition
  - statistical operations.

5. Data Output & Interpretation.
- The processed data is made readable & usable through:
  Graphs, reports, Text summaries

6. Data storage:-
Final results are stored in:
- Databases, cloud storage, files
- must comply with data protection laws like GDPR.

Types of Data processing.

1. Batch processing
Processes data in chunks at scheduled times
ex:- Monthly salary processing.

2. Real-time processing
Instant data processing as it arrives.
e.g:- fraud detection in banks.

3. Multiprocessing
use multiple CPUs to process data in parallel
e.g Big Data analytics.

4. manual processing
Done by humans: slower used where precision is vital.
e.g financial audit.

Challenges in data processing:

Organizations face several challenges when managing large volumes of data, including:
- Quality issues
- scalability constraints
- Integration complexity
- Regulatory compliance.

3. Explain in detail about data preprocessing with an example.

Data preprocessing is the process of preparing raw data for analysis by cleaning and transforming it into a usable format.

- Goal is to improve the quality of the data
- Helps in handling missing values, removing duplicates and normalizing data.
- Ensures the accuracy and consistency of the dataset.

Steps in Data Preprocessing:-

1. Data cleaning
2. Data integration
3. Data Transformation
4. Data Reduction.

1. Data cleaning

It is the process of Identifying and correcting errors or inconsistencies in the dataset. It involves handling missing values, removing duplicates, and correcting incorrect or outlier data to ensure the dataset accurate and reliable.

a. missing values:-

- place of missing values, we can replace with "NA"
- sometimes replaced with most probable values
- missing values can be filled in 2 ways

manual / automatic

b. Noisy Data

noisy data → inconsistent / error data

methods to handle

1. Binning method :-

The data is sorted into equal segments, and each segment is smoothed by replacing values with the mean or boundary values.

2. Regression

Numerical predication of data

3. clustering

similar data items are grouped at one place

2. Data Integration

merging of multiple heterogeneous source of data are combined into single dataset

-2 types of data Integration

a. Tight coupling

Data is combined together into a physical location.

b. loose coupling

Only an interface is created and data is combined through the interface and also accessed through interface

- Data remains in actual database only

3. Data Reduction

volume of data is reduced to make analysis easier

methods :-

a. Dimensionality Reduction :-

reduces no of I/p variables in the dataset bcoz, large I/p variables → poor performance

c. Data cube Aggregation
Data is combined to construct a datacube

c. Data compression:-
Reducing the size of data by encoding it in a more compact form, making it easier to store & process.

d. Numerosity Reduction:-
Reducing the no. data points by methods like sampling.

4. Data Transformation
Data is transformed into appropriate from suitable for mining process.

a. Data Normalization:
The process of scaling data to a common range to ensure consistency across variables.

b. Discretization:
converting continous data into discrete categories for easier analysis.

c. Data Aggregation:-
combining multiple data points into a summary form.

d. concept Hierarchy Generation:-
Organizing data into a hierarchy of concepts to provide a higher-level view for better understanding & analysis.

4. Explain about different tools used for DA. an example

Data analytics tools help in collecting, processing, analyzing, and visualizing data to gain insights.

some most commonly used tools

1. Excel

- Best for small-scale data analysis, reporting and visualization.

Ex:- A sales manager uses excel to analyze monthly sales trends and create pivot tables

2. Python

used for data manipulation, statistical analysis, and machine learning

Ex:- A data analysts uses Python pandas library to clean a dataset of customer transactions and create a predictive model for customer churn.

3. R

preffered for statistical computing & visualization

Ex:- A researcher uses R to analyze survey data and create regression models for predicting customer satisfaction

4. SQL

used for quering & managing large databases

Ex:- A bank uses SQL to extract customer transaction details & find activities.

5. Tableau
A powerful tool for interactive data visualization and dashboards.
ex:- A marketing team uses Tableau to create a real-time dashboard of customer engagement metrics.

6. Power BI
microsoft tool for business intelligence and data visualization.
ex:- A company's HR department uses power BI to track employee performance & attrition rates.

7. Apache Spark
Big data processing and real-time analytics.
ex:- A streaming platform like Netflix uses Apache spark to analyze user history

8. Hadoop:-
Storing & processing large-scale data across distributed systems.
ex:- An e-commerce site like Amazon use Hadoop

9. Google Analytics
website traffic analysis & digital marketing insights

10. KNIME
Data mining, machine learning & automation
Ex:- A pharmaceutical company uses KNIME

5. Explain Data modeling Techniques

Data modeling is the process of structuring and organizing data to make it useful for analysis.

Data models:-
These are visual representation of an enterprise's data elements and the connection blw them.

Types of Data models
3 types of Data models

1. Conceptual Data model
   - High level, abstract view of data
   - Focuses on business concepts and rules
   - used in the early project stage.
2. Logical Data model
   - Defines tables, columns & relationships.
   - Independent of any Database system
   - Basis for database design
3. Physical Data model
   - Actual database implementation
   - Includes tables, constraints & keys
   - Specific to database system (e.g. MySQL, Post

Types of Data modeling

1. Hierarchical model:
   - Data is structured like a tree (one parent, multiple childre
   - used less frequently today.
   - Ex:- A college (parent) has student & teacha (child)

2. Relational model :-
- Represents data in tables (rows & cols)
- Used in Relational database management system

3. Object-Oriented Data model.
- Data is stored as objects like in OOP
- supports abstraction, inheritance & encapsulation

4. Network model
- Uses a graph structure with nodes (data) & edges (relationships)
- supports multiple parent-child relationships.

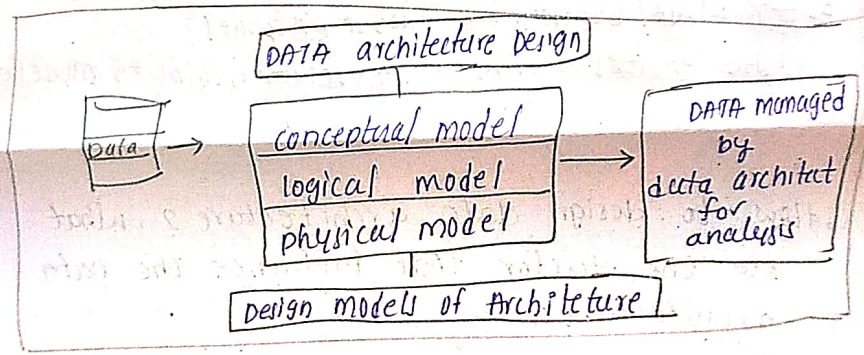
5. ER Model (Entity-Relationship model)
- Use entities, attributes & relationships to model data.
- Visualized with ER diagrams.

1. How to design data architecture? What are the factor that influence the data architecture?

- Data architecture design is set of standards which are composed of certain polices, rules, models and standards.
- which manages, what type of data is collected, from where it is collected, where storing that data for further analysis.
- Data architecture design is important for creating a vision of interactions occurring between data systems.

for example,

- If data architect wants to implement da architecture the visionary model of data interaction during the process can be achieved.
- Data architecture also describes the type of data structure applied to manage data and it provides as easy way for data preprocessing the data architecture is formed by dividing into 3 essential models & then are combined.



i. conceptual model

It is a business model which uses ER [entity relationship] model for relation blw entities and their attributes.

ii. Logical model:

It is a model where problems are represented in the form of logic such as rows, colm, of data, class, xml tags & other DBMS techn

iii. physical model:

physical models holds the data base design like which type of data base technology

Data
ll be suitable for suitable for architecture.

Factors that influence data architecture

1. Business Goals :-
   Determines the data flow & structure based on use cases
2. Data Volume & variety :-
   Affect storage & processing decisions
3. Scalability Needs
   Determines cloud vs on-premise architecture
4. Security & compliance.
   Requires encryption, access control, and legal compliance
5. performance & Latency :-
   Real-time processing needs vs batch processing
6. Integration Requirements
7. Cost & Budget.

UNIT - II

Data Analytics

- Introduction to Analytis
- Introduction to Tools and Environment
- Application of modeling in Business,
- ~~Database~~s
- Databases & Types of Data and variables
- Data modeling techniques.
- Missing Imputations etc
- Need for Business Modelings.

Data Analytics :-

- Data Analytics is a systematic process of converting raw data into valuable insights using tools, technologies and statistical techniques.
- These tools are used for ~~cleaning~~ collecting, cleaning, transformation and modeling.
- It supports decision-making and is crucial for business growth, risk management, and enhancing customer experience.
- Its applications span finance, agriculture, banking, retail, government and more.

Process of Data Analytics.

Data can This process is carried out by data scientists, data analysts and engineers using data pipelines.

1. Data collection :-
   Gathering raw data from various sources or subsets.
2. Data cleaning :-
   correcting data quality issues like errors and duplicates and white spaces which need to be corrected before moving to the next step.
3. Data Analysis and Interpretation :-
   creating and testing analytical models using

tools like Python, Excel, R, Scala, and SQL

4. Data visualization:-
Using charts and graphs to identify patterns and insights.

Types of Data Analytics:-

1. Descriptive performance: Summarizes past data to identify patterns and issuses.
-Used to compare past performance.
2. Real-time Analytics:- Analyzes data as it is entered to track trends and competitors.
3. Diagonstic Analytics:- Analyzes past data to find causes of anomalies using techniques like regression and correlation.
4. Predictive Analytics:- uses current data and machine learning to predict future outcomes.
5. prescriptive Analytics:-
Suggest best solutions for decision-making and automation.
-used in loan approvals, scheduling, pricing models.

Methods of Data Analytics

1. Qualitative Data Analytics:-
Analyzes non-numerical data like words, images and behaviours.

Some common qualitative methods are

-Narrative Analytics is used for working with data acquired from diaries, interviews and so on.
- Content analytics is used for analysis of verbal data and behaviours.
- Grounded theory is used to explain some given event by studying.

2. Quantitative Data Analytics
uses numerical data for statistical analysis.
-Hypothesis testing asses the given hypothesis of the data set.
- sample size determination is a method of taking a small sample from a large group of people and then analysing it.
-Average or mean of a subject is dividing the sum total no.s in the list by the number of items present in that list.

Skills required to Data Analytics
There are multiple skills which are required to be a Data analyst.
- Programming languages: R, Python.
- Database language: SQL
- Knowledge of machine learning, probability statistics, Data management, and Data visualization.

Data analytics jobs :-

- Entry-level roles
  - Junior Data Analyst
  - Junior Data scientist
  - Associate Data Analyst.
- Experienced-level roles.
  - Data Analyst
  - Data Enginner
  - Data Architect
  - Data scientist
  - Marketing Analyst
  - Business Analyst.

Introduction to Analytics

Analytics is a systematic process of collecting, processing, and analyzing data to discover meaningful patterns, insights and trends. It involves using various tools, technologies and methodologies to transform raw data into actionable information that supports decision-making.

In essence, analytics empowers businesses and individuals to make informed decisions by leveraging data-driven insights.

And types are same as Data Analytics types.

Introduction to Tools and Environment

Analytics environment :-

A Data Analytics Environment refers to the complete ecosystem that enables data collection, processing, modeling, analysis and visualization.

It includes:

- Data storage systems
- Data processing platforms
- programming environments
- Visualization tools
- Cloud and big data platforms.

This environment enables analysts to work with data at various stages - from acquistion and cleaning to modeling and presentation, and also helps data professionals carry out various analytical tasks efficiently and accurate.

Categories of Data Analytics Tools.

1. Programming languages
   Used to write scripts, automate tasks, build model, and Process data.
   Ex:- Python, R, SQL, Java.
2. Data storage tools :
   These tools are used to store, retrive and manage large volumes of data in structured or unstructured formats.
   e.g:- MySQL, postgreSQL for relational data and MongoDB, Hadoop & HDFS for non-relational & big data storage

3. Data processing tools

These tools help in cleaning, transforming & preparing data for analysis.

ex:- Python, R and pandas, spark, Excel etc

4. Visualization tools

These tools converted processed data into visual forms like charts, graphs and dashboards.

Tools are :- Tableau, Power BI, Matplotlib and seaborn.

5. Statistical and Machine learning tools

These are used to perform statistical tests, build predictive models, and discover patterns in data.

ex:- scikit-learn, TensorFlow, SAS and IBM SPSS

6. Big Data tools

These tools are designed to process, store and analyze very large datasets distributed across clusters.

Apache Hadoop handles batch processing, while Apache spark enables faster inmemory computation

7. Cloud platforms:-

They provides scalable computing resources and services for storing and analyzing data.

Ex. AWS, Google Cloude & Microsoft Azure.

In DA, tools and environment form the foundation for effective data handling. A well-structured analytics environment ensures scalability flexibility and accuracy in analytical tasks.

Application of modeling in Business

It refers to the use of mathematical, statistical and computational techniques to represent real-world business process and scenarios.

These models enable businesses to simulate outcomes, predict trends, and optimize decision-making across various departments such as sales, finance, marketing, operations & HR.

- Modeling supports data-driven decisions, helping organizations reduce risks, increase efficiency, and maintain and competitive edge in the market.

major Applications :-

1. Sales Forecasting

- predictive models are used to analyze historical sales data.
- Helps in forecasting demand, managing inventory and planning sales targets.
- Enables organizations to avoid overproduction or stockouts.

2. Customer segmentation.

- models like clustering and classification group customers by demographics, preferences or behaviour.

- Supports targeted marketing, personalization and improves customer retention.

3. Risk management
- Statistical and machine learning models assess financial credit, and operational risks.
- Used in credit scoring, fraud detection & insurance underwriting.
- minimizes losses by identying high-risk activities.

4. Supply chain optimization
- Models like linear programming and simulation optimize logistics, inventory and supplier decisions.
- Helps reduce costs, shorten delivery items & improve efficiency.

5. Financial planning & Budgeting.
- Forecasting and scenario analysis help in planning revenues, expenditures and investments.
- models simulate best-case, worst case and average case financial scenarios
- Supports sustainable business growth.

6. Product pricing
- Regression models and elasticity analysis determine pricing based on custmer demand and competitor strategies
- Helps balance profit margins with affordability compatitive pricing.

7. Market Basket Analysis
- Association rule mining identifies product purchase patterns
- Used in cross-selling, promotions, and retail shelf planning.
- Improves sales strategies by understanding customer buying behaviour.

8. Project Management
- Time series and resource allocation models help in planning timelines, budget, and resource use.
- Redues delays and ensures efficient project delivery.

Modeling in business is a powerful approach that supports strategic planning, problem solving, and performance optimization.

Databases & Types of data and variables

A database is a structured collection of data stored electronically. It allows easy access, management, and updating of data using Database Management System (DBMS) such as MySQL, Oracle, MongoDB, etc.

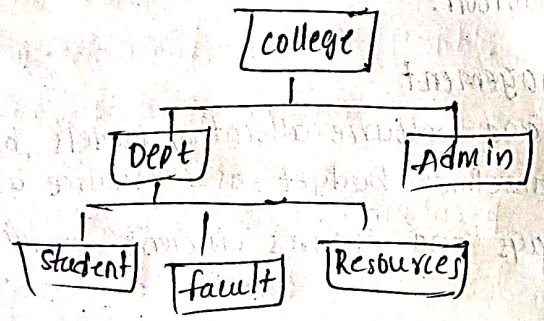
Characteristics:-
- Data organization - logical arrangement for quick access and modification.
- Efficiency - Optimized performance for storage & retrival
- Scalability :- Can handle large, growing volumes of data.
- Security :- ensures safe access & privacy.

Types:-

Databases can be classified based on their structure, usage, storage methods, and intended applications.

1. Hierarchical Databases

Tree-structured parent-child model



Ex:- IBM IMS

Adv:- Simple and fast for structured data

disadv:- Rigid structure, limited flexibility.

2. Network Database

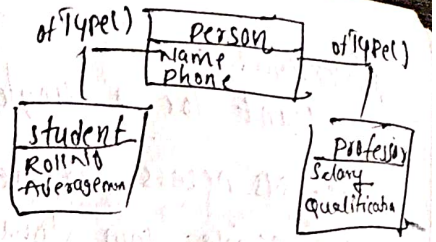
A Network databases builds on the hierarchical model but allows child records to link to multiple parent records, creating a web-like structure of interconnected data.

Ex:- IDS - Integrated Data Store

Adv:- 1. More flexible than hierarchical models
2. Handles complex relationships.

Dis:- Hard to design / manage

2. Object-oriented



- Data stored as objects with attributes and methods.

Ex:- Berkeley DB

Adv:- Good for multimedia reusability

dis:- Requires OOP knowledge.

4. Relational

Data in tables with rows (records) & columns (attributes)

Ex:- MySQL, PostgreSQL

Adv:- Simple, ACID (Atomicity, consistency, Isolation Durability)

dis:- Difficult to scale, schema-dependent.

5. Cloud:-

A cloud database operates in a virtual environment hosted on cloud computing platforms. It,

Ex:- AWS, Azure

Adv:- pay-as-you-go flexible

dis:- Internet dependence Security risks.

6. Centralized:-

It is a database stored and managed at a single location, such as a central server or data center.

Ex:- Central servers

Adv:- Easy security control

dis:- Limited performance at scale.

7. Personal :-
Meant for a single user lightweight
ex:- MS Access, SQ Lite.
Adv:- Simple, low storage.
disad:- Not scalable, fewer features

8. NO SQL :-
Non-relational, supports flexible structures
(JSON, key value)
Ex:- MOngo DB
ad :- Fast, scalable, unstructured
dis :- limited GUI, weak backup tools.

Types of Data in Analaytics

- Structured Data
Organized in a fixed format, usually in tables
ex:- Excel files, SQL databases.

- Unstructured data
No predefined format. Requires more processing
ex:- Emails, PDFs, videos.

- Semi-structured Data.
partially organized data with tags or markers
ex:- JSON, XML

- Time Series Data:
Data points indexed in time order
ex:- stock prices, temperature logs.

- Spatial Data :- ex: Maps, GPS Data
Data that represents locations & shapes.

Types of Variables:-
variables represent measurable characteristics or properties of data. They are key to statistical and machine learning analysis.

- Nominal
categories without any order
ex:- Gender (Male, Female), colors.

- Ordinal
categories with a meaningful order
eduction level (High school, B E)

- Interval :- Numerical data without a true zero.
e.g : Temperature in Celsius.

- Ratio: Numerical data with true zero.
Age, Height, weight, sales.

Classification of Variables

a. Quantitative (Numerical) Variables.
These represent measurable quantites (countable)
- Discrete Variable: can take only specific values
ex:- No. of students, cars.
- continous Variables:-
can take any value within a range (measurable)
eg :- Height, weight, temperature.

b. Qualitative (categorical) Variables:-
These represent categories or labels
- Nominal: No order
- Ordinal: Ordered categories.

- Databases form the backbone of modern data storage and management
- Understanding the types of data (structured, unstructured) and types of variables (categorical, numerical) is crucial for effective data analysis.
- Proper classification helps in choosing the right tools, techniques and visualization methods in data analytics.

Data modeling techniques

Data modeling:-

It is the process of structuring, organizing and representing data to define how it is stored, connected and accessed in databases or analytics systems.

It helps analysts and businesses understand:
- what data is collected.
- How data elements relate
- How to efficiently store and retrive data.

Modeling techniques

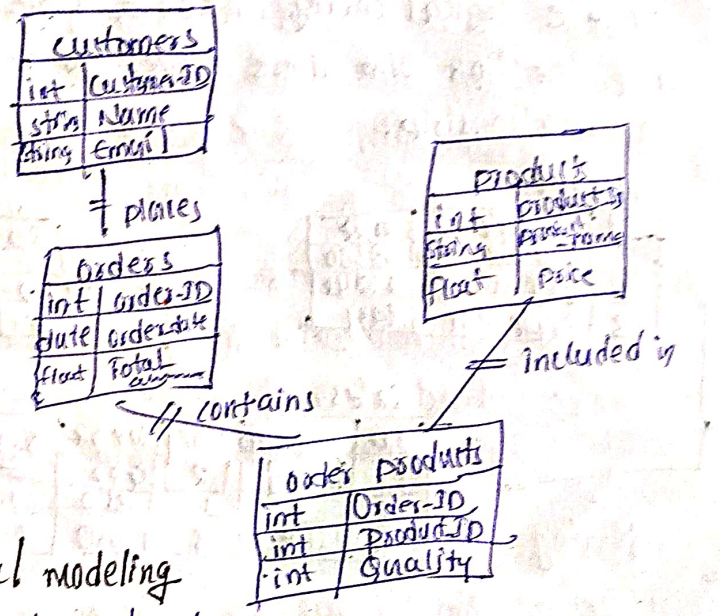
1. Entity - Relationship (ER) Modeling

It is one of the most common techniques used to represent data. It's concerned with defining 3 key elements.
- entites (objects or things within the system)
- Relationships (how these entities interact with each other)
- Attributes (properties of the entities)

The ER model provides a clear, visual representation of how data is structured to help map the connections b/w different data points.

Ex:-

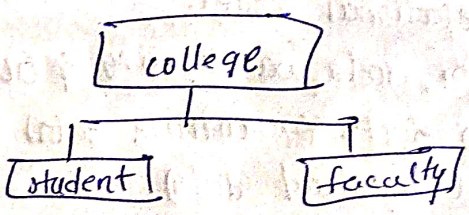
E-commerce store.



2. Hierarchical modeling

It is like a tree structure

The data is stored in parent and child records, each of which may have a set of attributes



Limitation:-

Hard to manage many-to-many relationships.

3. Relational Modeling

- Stores data in tables (relations) with rows and columns
- Based on keys (primary, foreign) and constraints.
- Supports SQL-based systems like MySQL, PostgreSQL

Each table contains information relevant to a singale logical entity, like a store for example and the link blw these tables is represented by a relationship.

Product

Id	Name	Price
1	washingMach	471.00
2	4K flatScreen	1250.4K
3	Apple MBAC	999

Store

ID	Name
1	Electronic outlet
2	Future Tech
3	Quatum Tks

Sales

ID	ProductID	StoreID	SalePrice	Quantity
1	1	2	472	2
2	2	1	1350	1
3	3	3	1050	1

4. Dimensional modeling

The dimensional data model is allso a variation of the relational model.

- Designed for data wasehousing & OLAP systems
- focuses on facts (quantitive data) and dimensions (contextual data)
- used to build star schema & snowflake scheme

5. Object-oriented modeling

- Represents data using objects
- supports inheritance, encapsulation & polymorphism
- Used in object-oriented databases and modern programming.

6. Conceptual, logical & physical models

- Conceptual :- High ~~focuses~~ level, focuses on what data exists and its relationships (ER diagrams)
- logical :- Defines how data is organized logically (tables, fields)
- physical - specifies how data is stored physically on system (indexes, partitions)

Using appropriate data modeling techniques ensures efficiency, clarity and scalability in both operational and analytical systems. It acts as a bridge blw real-world data needs and technical implementation.

I'm Missing Imputations

what is missing Data. ?

Missing data refers to the absence of a data value in a dataset where it should exist.

Impacts :- • leads to inaccurate analysis & biased result
• Affects model performance in ML
• Can cause misleading interpretation.

Types of missing data.

1. MCAR (missing completely at random)
   missing values occur independently of any data values
   ex:- A survey respondent accidentally skips a question

2. MAR (Missing at Random)
   Missingness is related to observed data but not the missing value itself.
   ex:- People with higher income levels tend to skip the salary question.

3. MNAR (missing Not a Random)
   missing value depend on the variable itself
   ex:- Individuals with very low income purposely leave the income field blank.

Missing Imputation:-

It is the process of filling in or replacing missing values in a dataset to ensure that the data is complete and usable for analysis or modeling.

-Imputation techniques

1. Deletion methods:
   - listwise Deletion:
     Remove the entire row if any value is missing
   - Pairwise Deletion:
     Uses all avaiable data without removing the row entirely useful in correction analysis.

2. mean / median / mode Imputation
   - Replace missing values with the mean, median or mode
   - Easy but may reduce variance and introduce bias e.g:- student age

A	20	mean = (20+22)/2
B	22	= 21
C		B = 21

3. forward / Backward fill:-
   → forward fill (ffill): use the last known value
   → backward fill (bfill): use the next known value.

4. k-Nearest Neighbors (kNN) imputation
   Missing values are filled using the average value of the k most similar rows.
   e.g:- If a row is missing age, KNN find k rows with similar income and education and averages their ages.

5. Regression Imputation
   Predicts the missing value using regression models based on other variables.
   ex:- Predict missing income using age and education.

6. Multiple Imputation:-
   - Generates multiple datasets with estimated values and averages the result.
   e.g:- Useful when the dataset has 20-30% missing value, & uncertainty must be preserved.

Missing imputations are essential for preparing clean and usable data. Depending on the nature of the data & missingness, the appropriate technique must be selected to preserve data integrity, reduce bias, and improve model accuracy.

Need for business modeling

what is business modeling

Business modeling refers to the process of creating abstract representation of business processes, operations and data flows using mathematical, statistical, or computational tools.

These models help stakeholders to ~~analyze~~ understand, analyze and improve business performance and act as the foundation for effective decision-making, strategic planning, and IT alignment.

Why is needed.

1. Clarity in business processes
   - Provides a visual representation of workflows & activity
   - Helps stakeholders understand process interconnections.
   - Aids in identifying areas for process improvement.
2. Improved Decision-making
   - Offers data-driven insights through simulations and "what-if" analysis.
   - Helps managers take informed decisions backed by analytical models.
3. Alignment blw Business & IT
   - Acts as a bridge blw business teams & technical developers.
   - Ensures that IT solns align with organizational goals & workflows.
4. Process Optimization and efficiency.
   - Identifies redundancies, bottlenecks, & inefficiences in workflows.
   - facilitates workflow redesign to improve productive & reduce costs.
5. Strategic planning & forecasting
   - models break down complex problems into manageable components.
   - Used in "what-if" analyses to test outcomes for different strategies.
6. Communication & Documentation:
   - A business model acts as a common language across departments.
   - Helps in on boarding, training & maintaing consistency in operations.

7. Improved customer understanding
- models analyze customer behaviours, preferences and feedback.
- enables targeted marketing, product recommendation and better user experience.
e.g:- E-commerce platforms model customer buying patterns.

8. Competitive Advantage:-
- modeling helps businesses adapt faster, make smarter moves, & stay ahead of competitors
- Transforms raw data into strategic insights.

Real-time example:-
Amazon - E commerce & Marketplace model

Business modeling concepts used:-
- process optimization:- Amazon uses logistics & inventory models to streamline its global supply chain
- Customer segmentation
- Revenue simulation:-
models forecast earning from Subscriptions Vs 3rd party sales.

Why it fits:-
Amazon's model is a perfect case of multi channel business modeling - blending retail, platform, logistics & subscriptions, all aligned via structured data models for decision making & forecasting.

Business modeling is essential for organizations to analyze operations, optimize workflows and innovate with confidence. It improves communication, enables strategic planning and drives data-backed decision-making.

In a competitive environment businesses that leverage models gain a clear advantage in term of efficiency, adaptability & customer satisfication.

Unit-III

Regression

- Concepts, Blue property assumptions,
- Least square Estimation
- variable Rationalization
- ~~Exam~~ Model Building etc.

Logistic Regression:-

- Model Theory
- Model fit statistics
- Model construction
- Analytics applications to various Business Domains etc.

Regression analysis

- It is a statistical technique used to examine the relationship between dependent and independent variables.
- It determines how changes in the independent variable(s) influence the dependent variable, helping to predict outcomes, identify trends and evaluate causal relationships.
- widely used in fields like business, economics, healthcare and social sciences, regression analysis provides a robust framework for data-driven decision-making.

It helps in

- predicating future values
- understanding variable relationships.
- Quantifying the impact of predictors.

Methods of Regression Analysis

1. Ordinary least squares (OLS)

It minimizes the sum of squared errors between actual and predicted values

$$\text{Minimize } \sum (y_i - \hat{y}_i)^2$$

Use case:-

- most commonly used method in simple and multiple linear regression.
- Suitable when assumptions like linearity & homoscedasticity are met.

2. Maximum likelihood estimation(MLE)

- It choose parameter values that maximize the likelihood that the observed data occured

$$\hat{\theta} = \arg\max_{\theta} L(\theta|x)$$

Use case:-

commonly in logistic regression and generalized linear models.

3. Ridge and Lasso Regression.

Both are regularization techniques

- Ridge adds L2 penalty → shrinks all coefficients
- Lasso adds L1 penalty → can set some coefficients to zero (feature selection)

Ridge : $\Sigma(y_i - \hat{y}_i)^2 + \Sigma \beta_j^2$

Lasso : $\Sigma y_i - \hat{y}_i)^2 + \lambda \Sigma |\beta_j|$

use case:-

- Useful when dealing with multicollinearity or high-dimensional data.
- prevents overfitting.

4. Stepwise Regression:-

Adds or removes predictor systematically to improve the model.

Use case:-

Used in exploratory data analysis to identify significant ~~ped~~ predictors.

5. Bayesian Regression:-

Incorporates prior beliefs and data evidence into regression modeling

Use case:- forecasting

Useful in ~~exploratory~~ data analysis to identify significant predictors.

Types of Regression Analysis:-

1. Simple linear Regression

Assume that there is only one independent variable x. If the relationship b/w x (independent variable) and y (dependent or output variable) is modeled by the relation,

$y = a + bx$

then the regression model is called a linear regression model.

Use case:- predicting housing prices based on square footage

2. multiple regression

Assume that there are multiple independent variables say $x_1, x_2 \ldots x_n$. If the relationship one dependent and 2 or more independ variables

$y = a_0 + a_1 * x_1 + a_2 * x_2 + \ldots a_n * x_n$

Use :- Analyzing how advertising budget, product prize and seasonality affect sales revenue.

3. Logistic Regression:-

Used when the dependent variable is binary e.g yes/no, success (failure). $P(y=1) = \frac{e^{a+bx}}{1+e^{a+bx}}$

Use:- Predicting wheather a customer will buy a product

based on demographic features.

4. Polynomial Regression.

Assumes that there is only one dependent variable x. If the relationship blw independent variables x and dependent or o/p variable y is modeled by the relationship

$$y = a_0 + a_1 * x + a_2 * x^2 + \dots a_z * x^n$$

for some positive integer $n \geq 1$, then we have a polynomial regression.

BLUE property of Estimators (Gauss-Markov Theorem):

The term BLUE stands for
Best linear unbiased estimator.

variance tells us how much the data is spread out.
variance = How far values are from the average(mean)

What is an estimator
In statistics an estimator is a rule for calculating an estimate of given quantity based on observed data.

ex:- x follows a normal distribution, but we don't know the parameters of our distribution, namely mean (μ) and variance (σ^2)

To estimate the unknowns the usual procedure is to draw a random sample of size 'n' and use the sample data to estimate parameters

OLS is a method used in linear regression to draw the best possible straight line through a set of data points.

- It helps us predict something based on something else.

OLS chooses the line that minimizes the error
It calculates the square of each error, adds them all, and finds the line where this total squared error is smallest.

Ordinary least squared = least total of squared errors.

What is BLUE
In linear regression, the term BLUE refers to the properties of an estimator that is
BEST LINEAR UNBIASED ESTIMATOR

These properties are guaranteed for ordinary least squares (OLS) estimators if certain assumptions are met. This is formalized in Gauss-Markov Theorem.

meaning of each term:-

Best – Has the minimum variance among all linear biased estimators.
Linear – Is a linear function of the observed data.
Unbaised – On average, the estimator gives the true value of the parameter.
Estimator – A rule or formula used to calculate an estimate from sample data.

Gauss-Markov Theorem

These states that :-

Under certain assumptions, the OLS estimator is Best Linear Unbiased Estimator (BLUE) of the regression coefficients.

Assumptions required for BLUE

1. Linearity :- The model is linear in parameters

$$y = \beta_0 + \beta_1 x + \varepsilon$$

2. Random sampling :- Data is collected via random sampling.

3. No perfect multicollinearity :-

Independent variables are not perfectly correlated.

4. Zero mean of errors :- $E[\varepsilon] = 0$, meaning errors average to zero.

5. Homoscedasticity :-

constant variance of error terms across observations.

6. No Autocorrelation :- Errors are not correlated with each other.

Why BLUE is important :-

- ensures that OLS estimates are efficient and accurate.
- Forms the foundation of linear regression analysis
- If assumptions are violated, estimates may still be unbiased but not best.

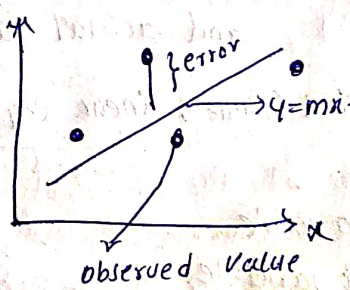
The BLUE property ensures that OLS regression estimators are statically optimal when the Gauss-Markov assumptions hold true. It gives confidence that prediction are reliable, unbiased and have minimum variance — a key strength in data modeling and analytics.

Least Square Estimation

It is a method used in linear regression to find the best fitting straight line through a set of data points.

It works by :-

minimizing the sum of squares of the errors (the differences blw actual values and predicted values).



least squares method is used to derive a generalized linear equation between two variables. when the value of the dependent and independent variables they are represented as x and y coordinates in 2D cartisian coordinate system.

The least squares method is a popular statistical technique used in
- Regression analysis - Data fitting
- predictive modeling.

Its main purpose is to find the line of best fit that minimizes the sum of squared differences between the actual data points and the values predicated by the line.

Why is it used?-

- To predict a dependent variable (y) from an independent variable (x)
- To draw a straight line that fits the data as closely as possible.
- To reduce the total error (difference blw predicated and actual values)

Basic equation of the line (simple linear Regression)

$$y = mx + c$$

where :- y : Dependent variable
x :- Independent variable.
m :- slope of the line
c :- y :- Intercept.

Least squares formulas

- slope (m)

$$m = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

Intercept (c):

$$c = \frac{\Sigma y - m(\Sigma x)}{n}$$

These formulas help us calculate the best fitting line that minimizes the sum of squared errors.

Steps in the least square method

1. Identify x and y (independent and dependent variables)
2. Create a scatter plot of the data points.
3. Calculate the mean of x and y
4. Use formulas to calculate the slope m and intercept c.

$$m = \frac{\Sigma(x - xi)(y - yi)}{\Sigma(x - xi)^2} \qquad c = y - mx$$

5. write the equation of the best fit line as

$$y = mx + c$$

Example :-

(1,3) (2,4) (4,8) (6,10) (8,15)

step 1 : find the mean x and y

$$x = \frac{1+2+4+6+8}{5} = 4.2$$

$$\bar{y} = \frac{3+4+8+10+15}{5} = 8$$

step 2 :- Use the formula

x_i	y_i	$X - x_i$	$Y - y_i$	$(X-x_i)(Y-y_i)$	$(X-x_i)^2$
1	3	3.2	5	16.0	10.24
2	4	2.2	4	8.8	4.84
4	8	0.2	0	0.0	0.04
6	10	-1.8	-2	3.6	3.24
8	15	-3.8	-7	26.6	14.44
				55.0	32.8

step 3 :- Calculate slope m

$$m = \frac{\Sigma(X-x_i)(Y-y_i)}{\Sigma(X-x_i)} = \frac{55}{32.8} = 1.68$$

step 4. Calculate intercept c

$$c = y - mx = 8 - (1.68 \times 4.2) = 0.94$$

Final line :-

$$y = 1.68x + 0.94$$

Graph Interpretation :-

- The red dots represent the actual data points
- The straight red line shows the best fit line
- This line can now be used to predict y values for any x.

Limitations :-

- Not suitable for non-linear data
- Gets affected by outliers
- Assumes equal variance and normal distribution of errors.

The least squares method helps you find the best line through data points by minimizing total squared error ~~between~~ between actual and predicted values. It is essential for linear regression and accurate prediction in business, science and engineering.

Variable Rationalization

It is the process of identifying, selecting and transforming variables in a dataset to ensure that only the most relevant and meaningful variables are used in the analytical model.

- Improve the accuracy of a model
- Reduce noise and redundancy
- make the model easier to understand and faster to compute.

Why it is Needed :-

When building a regression model or any analytical model:

- you may have too many variables
- some variables may be irrelevant, correlated or repetative
- keeping only the meaningful variables leads to better performance.

Steps in Variable Rationalization:-

1. Understanding Data context:-
   - Review variables based on domain knowledge.
   - Identify which variables are logically important.
2. Feature selection
   Choose only the most important variables for the model.
   methods:- • correlation analysis (remove highly correlated variables)
   • p value (in regression, if $p > 0.05$, variable may be dropped)
   • forward / Backward / Stepwise selection)
3. Feature transformation
   Transform variables to make them more useful.
   ex:- • Taking logarithms of large nos
   • Creating polynomial features for non-linear patterns.
   • Normalizing / scaling variable to same range.
4. Remove Redundant or Noisy variables.
   - Remove variable that:
   • Repeat the same information as others
   • Contain too many missing values.
   • Add randomness, but not meaning.
5. Domain-Based selection
   Sometimes, your knowledge of the business or domain helps decide which variables are meaningful and which are not.

Example:-

Imagine you're predicting student marks.
You collect these variables.
- Hours studied
- favourite color
- Age
- Mobile brand
- Attendence %

After rationalization, you keep
- Hours studied
- Attendence %

Remaining are not required.

Tools:-

- correlation matrix
- Heatmaps
- variance Inflation factor (VIF)
- PCA (principle component analysis)

Advantages:-

- Reduce time model complexity
- Improves prediction accuracy
- Faster computation
- Reduce overfitting

Dis:-

- can be time consuming
- Risk of removing useful variables
- may introduce bias
- Requires expertise

Summary:-

Variable Rationalization is the process of selecting the most relevant variables and removing or transforming those that are unnecessary or harmful to model accuracy.
It improves performance, simplifies models and helps make better predictions.

Model Building

It is the process of creating a predictive or analytical model using data. The goal is to understand relationships between variables or predict outcomes accurately.

Model building is the process of
- preparing data.
- choosing a model
- Training it to learn from the data.
- Evaluating its performance.
- And using it for predications or analysis.

The main goal is to create a model that can understand patterns in training data and predict accurately on new/unseen data.

Key objectives
- Achieve high accuracy on training & testing dataset
- Ensure the model generalizes well, not just memorizes data.
- make the model suitable for real-world use.

Steps in ~~do~~ model building:-

1. Split ~~the Dataset~~

To avoid overfitting, we split the data into:-
- Training dataset - used to train the model
- Testing dataset - used to evaluate the model's performance.

Usually done in a 75:25 or 80:20 ratio using python's train_test_split.

code:-
```
from sklearn.model_selection import train_test_split
x = data[['x', 'y']]
y = np.random.rand(1000)
train_x, test_x, train_y, test_y =
        train_test_split(x, y, test_size=0.21)
```

2. Scale the Data

Why?
- prevent features with big values from dominating small-scale feature.
- make models like KNN and Gradient descent accurate and faster.

→ use MinMaxScaler to scale data to range [0, 1]

```
from sklearn.preprocessing import MinMaxScaler
scaler = MinMaxScaler()
train_x_scaled = scaler.fit_transform(train_x)
train_x_scaled = scaler.transform(test_x)
```

3. Choose the Model

Depending on the problem

Regresion — linear regression. Decision Tree
classficaction — logistic regression, KNN, SVM.

here we are using regression problem so we use Decision tree.

```
from sklearn.tree import DecisionTreeRegression
reg = DecisionTreeRegressor(min_samples_split=4,
                            max_leaf_nodes=10)
reg.fit(train_x_scaled, train_y)
```

4. make predictions

5. Evaluate the model
   we measure how well the model predicated using Mean squared Error (MSE)
   - A lower MSE means better performance.

6. Visualize the model.

Advantages :-
- Easy to interpret
- can handle both numerical & categorical data
- visualize structure make decisions transparent.

limitations :- - can overfit if not pruned or limited
- sensitive to small changes in data.

Summary :-

Model building is the process of training a machine to learn patterns from data and make predictions. It includes data splitting scaling, training, testing, evaluation & visualization.

Logistic Regression

logistic regression is a statistical technique used to :
Predict the probability of a binary outcome (e.g Yes/NO, 0/1, True/false).

Unlike linear regression (which predicts a continuous value), logistic regression is used when the target variable is categorical.

Real-life example

- will a customer buy a product ? (Yes or No)
- Is the email spam ? (spam or Not spam)
- will a student pass the exam? (pass or fail)

Model theory of logistic regression

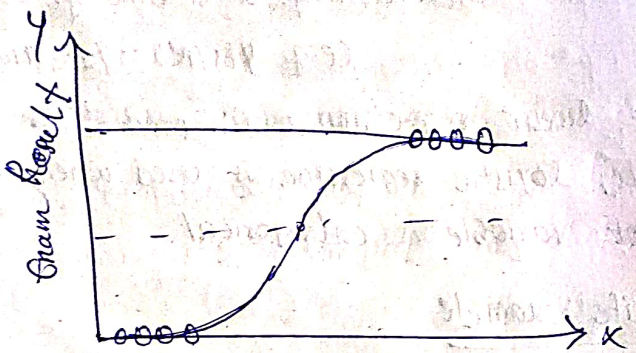
It is a classification technique used to predict the probability of a binary outcome. e.g (Yes/No)

- Why not use linear Regression :-
  It predits any value, even below 0 or above 1 which is not vaid for probability.
- Logistic Regression fix :-
  It uses the sigmoid function to keep the output between 0 and 1.
  Sigmoid function (S-shaped curve)

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \text{where } z = B_0 + B_1 x_1 + B_2 x_2 \ldots B_n x_n$$

Output: A probability between 0 and 1

- If $\sigma(z) > 0.5$, predict class 1
- If $\sigma(z) < 0.5$, predict class 0



- Logit function (Log-odds)

Logistic regression predicts the log of odds, which is a linear function.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

- Loss function: Binary cross-entropy

To train the model, logistic regression uses the Log loss or Binary cross-entropy function. It penalizes wrong predictions more harshly.

Ex:- predict if a student passes based on hours studied.

Hours (x)	passed (y)
1	0
2	0
3	0
4	1
5	1

If the trained model gives $\beta_0 = -4$, $\beta_1 = 1$

Then $z = -4 + (1 \times 3) = -1$

$$\sigma(-1) = \frac{1}{1+e^{-(-1)}} = 0.27 \Rightarrow \text{predict class 0}$$

For 5 hours:-

$$z = -4 + 1 \times 5 = 1$$

$$p(y=1 \mid x=5) = \frac{1}{e+e^{-1}} \approx 0.731$$

73.1% → predict pass (1) probability of passing.

Here β_0 & β_1 are not calculated manually like linear regression.

And in β_0 (intercept) and β_1 (slope). here they are use logit function.

Advantages:-

- Simple and easy to implement
- predict probability (0 to 1)
- Efficient for small datasets.
- less prone to overfitting.

Applications:-

- Healthcare - Banking & finance, - Email filtering
- Education.

- $\beta_1 > 0$: As x increases, probability of class 1 increases
- $\beta_1 < 0$: As x increases, probability of class 1 decreases
- Exponentiating β_j gives the odss ratio:

Odds Ratio $= e^{\beta_j}$

Ex:- • $\beta_1 = 1 \Rightarrow e^1 \approx 2.718$

→ extra hour of study doubles the odds of passing

Model fit statistics

model fit statistics help us evaluate how well a logistic regression model represents the data i.e. how accurate, meaningful, and reliable it is.

1. Deviance

It is like the "error" in logistic regression

- Null Deviance: Error of a model with only the intercept (no predicators)
- Residual Deviance: Error after adding predicators (features)

Interpretation:-

- lower residual deviance = better fit
- compare both using Likelihood Ratio Test (LRT)

2. Likelihood Ratio Test (LRT)

It compares the null model and the full model

$$G = -2 \times (\log L_{null} - \log L_{full})$$

- If G is large, it means the model with predictors is significally better.
- Compare G to the Chi-square table for significance.

3. Pseudo R^2

Since regular R^2 (used in linear regression) doesn't apply - we use pseudo R^2 to measure how much variance is explained.

McFadden's R^2 $\quad 1 - \frac{\log L_{full}}{\log L_{null}}$

Cox & Snell similar to R^2 but not scaled to 1

4. AIC (Akaike Information Criterion)

$$AIC = 2k - 2 \log L$$

- k = no. of model parameters
- L : likelihood of the model
- lower AIC = better model

used for compering multiple models.

5. Confusion matrix.

A table showing the actual vs predicted values

Actual / pred	pred 0	pred 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy = TP+TN / (FP+FN TP+TN)

6. Precision and Recall

Precision:- How many predicited positives are actually correct.

$$\text{Precission} = \frac{TP}{TP + FP}$$

- Recall (sensitivity): How many actual positives are correctly predicited.

$$\text{Recall} = \frac{TP}{TP + FN}$$

7. F_1 score

It is the harmonic mean of precision and recall. It balances the two when you need a single performance measure.

$$F_1 \text{ score} = 2 \times \frac{\text{precision} \times \text{Recall}}{\text{precision} + \text{recall}}$$

8. ROC curve (Reciver operating characteristic)

plots ~~True~~ positive Rate (Recall) vs false positive Rate.

$$\text{false positive rate} = \frac{FP}{(FP+TN)}$$

9. AUC (Area under curve)

measures the model ability to distinguish blw classes.

1.0 - perfect model, 0.5 - No better than random

<0.5 worse the random.

Model construction:-

Assumptions:-

- Binary dependent variable (0 or 1)
- linearity the logit for continous variables
- No multicollinearity
- Independence of observations.

logistic function

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_n x_n)}}$$

Steps in construction:-

1. Define the problem
2. Select feature (x) and target (Y)
3. split data into training / testing.
4. Check multicollinearity
5. Train the logistic regression model.
6. Predict & evaluate. using accuracy, AUC etc.

Detailed flow of logistic Regression.

Input features $(x_1, x_2, \ldots, x_n)$ → linear combination $z = \beta_0 + \beta_1 x_1 + \cdots \beta_n x_n$ → sigmoid fn $\sigma(z) = 1/(1+e^{-z})$

↓

output probability $P(Y=1|X)$

↓

Thresholding

if $P > 0.5 \rightarrow 1$ else 0

↓

final classification

0 or 1

Analytics Applications to various Business Domains.

Data Analytics is the process of examining large datasets to discover patterns, trends and insights. It is widely used across different business domains to improve decision-making, enhance customer experience, reduce costs, and predict future outcomes.

1. Healthcare
   Predictive analytics for patient health monitoring disease prediction and hospital resource management
   e.g:- Using machine learning to predict patient readmission rates and optimizes hospital bed allocation.
2. Retail - E commerce
   Customer segmentation, demand forecasting and recommendation systems.
   e.g:- Amazon uses customer behavior data to recommend personalized products and optimize inventory.
3. Banking and finance
   Credit Risk modeling, fraud detection, customer segmentation.
   ex:- Banks use analytics to detect suspicious transactions in real-time to prevent fraud.
4. Manufacturing
   predictive maintenance, quality control, and chain optimization.
   e.g:- An automobile company uses sensor data from machines to predict equipment faciures before they occur.
5. Telecommunications :-
   Churn analysis, network optimization, and pricing strategies.
   e.g:- Telecome companies analyze call data to detect customers likely to switch to competitors and provide offers to retain them.
6. Logistics and supply chain.
   Route optimization, demand forecasting, and warehouse management.
7. Marketing and sales.
   campaign performance analysis, customer lifetime value prediction and ROI measurement.
8. Education
   student performance prediction, adaptive learning and resource allocation.

Analytics has become an essential tool in modern business operations. By converting raw data into actionable insights, organizations across domains can enhance decision-making, reduce cost

Unit – IV

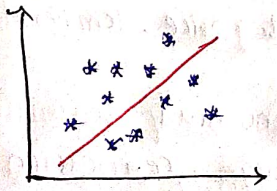
Object segmentation:

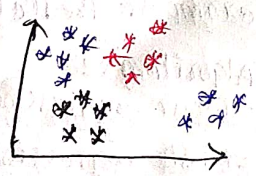
- Regression vs Segmentation
- supervised and Unsupervised learning
- Tree Building – Regression
- Classification
- Overfitting
- pruning and complexity
- Multiple decision Trees etc.

Time series method:

- Arima
- Measures of forecast Accuracy
- STL approach
- Extract features from generated model as Hieght, average, energy etc.
- ~~Average energy etc and~~ → Analyze for prediction.

aspect	Regression	segmentation
definition	Regression is a supervised learning technique used to predict a continuous numeric value based on input feature.	Segmentation (clustering) is an unsupervised learning technique used to group similar data points into clusters.
learning Type	Supervised learning	unsupervised learning.
objective	To predict a dependent variable (e.g sales, price, temperature.	To group similar data points without predefined labels.
output	A numeric value (e.g., house price = ₹50 lakhs)	A cluster or segment label (e.g., Customer Group A, B, C)
examples	predicting house prices, student scores, sales forecasting.	Customer segmentation, market segmentation, image segmentation.
Algo	linear regression, Polynomial regression, Decision Tree regression	k-means clustering, Hierarchical clustering DBSCAN.
Use case in Business	Forecasting future trends or demand	Identifying customer groups for targeted marketing.





Supervised and unsupervised learning

What is ML (machine learning)

It is a branch of artifical intelligence where machines learn from data without being explicitly programmed. ML is broadly classified into :-

1. Supervised learning
2. Unsupervised learning.

Supervised learning

It is a machine learning approach where the model is trained using labeled data (i.e, each input has a known output). The model learns the mapping functions from input to outputs.

Key concepts :-

Labeled data - Data with input-output pairs (X & Y known)

Goal - predict output values for new inputs

Output types - Continuous (regression) or categorical (classification)

Types of supervised learning

1. Regression :-

It is a type of supervised learning that is used to predict continous values.

example :- Predicting house price, temperature etc.

Algorithms :-

- linear regression
- Polynomial regression
- Lasso regression
- Ridge regression.

2. Classification

classification is a type of supervised learning that is used to predict categorical values, such as whether a customer will churn or not, whether an email is spam or not or whether a medical image shows a tumor or not.

- classification algo learn a function that maps from the input features to probability distribution over the output classes.

Algorithms :- logistic regression, support vector machine (SVM)

- Decision tree, - Random forest
- Naive Bayes.

Applications :-

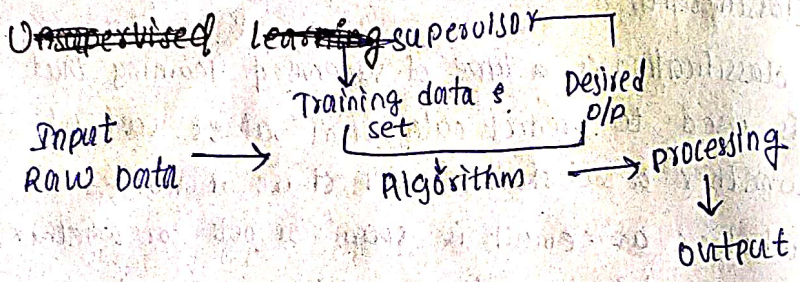
- Emails :- spam detection
- Heathcare :- Disease diagonis
- E-commerce :- product recommendation
- finance :- credit fraud detection
- NLP :- sentiment analysis translation.

Advantages

- Builds on previous experience
- Gets better with more data.
- solves real-world classification & prediction tasks
- Offers clear, interpretable results.

Disadvantages :-

- Needs large labeled datasets (costly to create)
- Can struggle with complex or unstructed data
- High computation and training time for large data.



Unsupervised learning

It is a machine learning method where the model is trained on unlabeled data. The goal is to discover patterns, structures, or relationships without any predefined output.

Key :-

Data - unlabeled (no output provided)

Goal - find clusters, associations or hidden structures

No supervision - No teacher or known answers during training.

Output :- Often groups, clusters or rules rather than specific predictions.

Input Raw Data → Interpretation → Algorithm → processing → output

Model training

Types of unsupervised learning :-

1. Clustering :-

Groups similar data points based on clusters based on distance (similarity).

clustering Algorithms work by iteratively moving data points closer to their cluster centers and further away from data points in other clusters.

Algo :- K-means clustering . Hierarchical clustering

- PCA (principal component analysis)
- DBSCAN
- Gaussian Mixture Models (GMM)

2. Association rule learning.

finds interesting rules between variables

e.g :- If a customer buys Bread, they are likely to buy Butter.

Algo :- Aprior Algorithm, Eclat Algorithm, Fb Growth.

Application :-

- E-commerce - customer segmentation for targeted marketing
- Recommendations - suggests movies / music / products.
- Science - finding hidden patterns in research data.
- Anomaly detection :- fraud or intrusion detection.

Advantages :-

- No need for labeled data
- Useful for data exploration
- Can reveal unknown patterns in complex datasets.
- Handles large and high-dimensional data well.

Disadvantages :-

No predefined o/p → Hard to measure accuracy

- Results may need manual validation
- Sensitive to outliers, noise and missing values
- may produce less precise groupings than supervised learning.

TREE BUILDING - Regression, classification

Tree based models are powerful tools for both regression and classification tasks. These models split data into smaller and smaller groups based on feature values, creating tree-like structure.

Tree building is the process of constructing a decision tree, which uses a series of decisions (nodes) to predict an output. It splits dataset into branches based on feature values until a final decision (leaf) is made.

CART (Classification And Regression Trees) is a variation of the decision tree algorithm. It can handle both classification and regression tasks.

CART is a predictive algorithm used in machine learning and it explains how the target variables values can predicted based on other matters,
- It is a decision tree where each fork is split into a predicator variable and each node has a predication for the target variable at the end.

- The term CART serves as a generic term for the following categories of decision trees:

- Classification trees:- The tree is used to determine which "class" the target variable is most likely to fall into when it is continuous.

- Regression trees:-
  These are used to predict a continuous variable value.

CART Algorithm:-
It is a decision tree algorithm that is used for both classification and regression tasks. It is a supervised learning algorithm that learns from labelled data to predict unseen data.

- Tree structure:- CART builds a tree like structure consisting of nodes and branches. The nodes represent different decision points, and the branches represent the possible outcomes of those decisions.
- Splitting criteria:-
  CART uses a greedy approach to split the data at each node.
  For classification tasks, CART uses Gini impurity as splitting criterion. For regression tasks CART uses residual reduction as splitting creterion.
- Pruning:- To prevent overfitting of the data, pruning is a technique used to remove the nodes that contibute little to the model accuracy.

How does CART work

1. Try all possible splits for each input variable
2. pick the best split (based on Gini impurity or error reduction)
3. Split the data into two parts using the ~~split~~ best split.
4. Repeat for each new part (sub-tree) until:
   - Tree gets too deep
   - Or there's no better split left.

What is Gini Index / Gini impurity?

- The Gini Index tells us how impure or mixed a group (subset) is.
- It is used in classification problems to decide where to split the data in a decision tree

Gini impurity checks: "what is the chance of making a wrong guess if we randomly assign a label from a group?

Formulas- $Gini = 1 - \sum_{i=1}^{n} p_i^2$ p_i – probability, n – no. of class.

Gini Index values-
- 0 → perfectly pure (only one class present)
- Close to 1 → highly impure (many mixed classes)
- 0.5 → Two classes with equal 50-50 probability.

ex:- Imagine you have a node with
- 4 apples
- 6 oranges

$p_{apple} = 0.4$, $p_{orange} = 0.6$

$Gini = 1 - (0.4^2 + 0.6^2) = 1 - (0.16 + 0.36) = 0.48$

This node is not pure because it has a mix of 2 classes.

Why CART uses Gini?
- CART looks for the split the minimizes the Gini impurity.
- At each node, it tries all possible splits, calculates the Gini for each, & pick the best (lowest Gini)
- Lower Gini = more pure = Better split.

CART for classification

In classification, CART is used when the output (target variable) is categorical – for example Yes/No, pass/fail, Spam/Not spam.

How it works
1. start with the full dataset (root node)
2. Check all feature
3. for each feature, try different split values.
4. Calculate Gini impurity for each possible split
5. choose the split that gives the lowest Gini impurity (most pure groups)
6. Repeat the process for each branch until
   - Nodes become pure
   - Tree reaches a maximum depth
   - There's not enough data to continue split.

CART for Regression

CART for regression is a decision tree learning method that creates a tree like structure to predict continuous target variables.

How does it works

It works by splitting the training data recursively into smaller subsets on specific criteria. The objective is to split the data in a way that minimizes the residual reduction in each subset.

- Residual Reduction:-

It is a measure of how much the average squared difference blw the predicted values and the actual values of the target variable is reduced by splitting the subset.

- Splitting Criteria:-

It evaluates every possible split at each node and selects the one that results in the greatest reduction of residual error in the resulting subsets.

CART model representation:-

It builds decision trees step by step based on data.

Here is how the CART model is formed.

1. Greedy Algorithm
2. Stopping criterion.
3. Tree pruning
4. Data preparation for CART.

Overfitting

It happen when a machine learning model learns the training data too well.

Instead of just learning the general pattern, it also learns noice, outliers, or random fluctuations.

*Think of it like a student memorizing answers instead of understanding concepts they score high in practice tests but fail in actual exam.

causes:- model is too complex
- Not enough training data
- Too many parameters compared to the data size.
- Training for too many iterations.

Effects of overfitting:-

on training data	on test data
• very accurate (low error)	poor performance (high error)
• looks perfect	fails to generalize.

example:- If you train a decision tree to keep splitting until each node has just 1 sample, it will:
- Give 100% accuracy on training data.
- But fail on new, unseen examples because it memorized, not learned.

How to detect overfitting

- Training accuracy is very high, but test accuracy is low.
- Performance changes a lot when switching bw datasets (i.e., high variance.
- Validation error starts increasing while train error keeps decreasing.

Simple Tricks to avoiding:-

- prune the decision tree
- Use cross-validation to monitor performance during training.
- limit model complexity.
- more training data.

Pruning

It is the process of removing unnecessary parts of a decision tree to make it simpler and less likely to overfit.

Why use pruning:-
- Reduces overfitting
- Improves model accuracy on test data.
- speeds up predictions
- makes the tree easier to understand.

Types of pruning

1. Pre-pruning (Early stopping)

Stop growing the tree before it becomes too complex.

- common stopping conditions
  - Tree reaches a maximum depth.
  - A node has too few samples
  - The gain (impurity reduction) from a split is too small.

* Saves time and prevents the tree from getting too large in the first place.

2. post-pruning (simplification After Growth)

first build the full tree, then cut back branches that are not useful.

* keeps only the splits that improve accuracy or reduce complexity.

This is often done using a validation set.

ex:-

Imagine a tree branch that classifies just 2 out of 1000 training samples.

If those 2 samples don't really change the models performance, we can prune that branch and avoid learning noise.

Pruning = "Trimming the decision tree to keep only the useful parts".

Complexity :-

Complexity refers to how large or detailed a decision tree becomes

This includes:-
- Tree Depth - How many levels it has
- Number of nodes
- Splits - How many times the data is divided.

Why does complexity matters:-

High complexity Tree:-

- learns detailed patterns and can fit complex patterns
- May overfit:- memorizes training data, including noise.

Low complexity Tree:-

- Too simple to learn actual parameters
- May underfit:- performs poorly even on training data

complexity control parameters:-

- Tree depth limit
- Minimum samples per leaf or split
- minimum impurity decreases required for split.

Goal:-
To balance complexity so the model is not simple or too complex - this helps in achieving good accuracy and better generalization to unseen data.

Multiple Decision Trees

Instead of using just one decision tree, we use many decision trees together
This helps:
- Reduce overfitting
- Increase accuracy
- Improve generalization

This concept is called ensemble learning.

Why use Multiple Trees

A single tree can:
- memorize training data (over-fit)
- Be sensitive to noise.

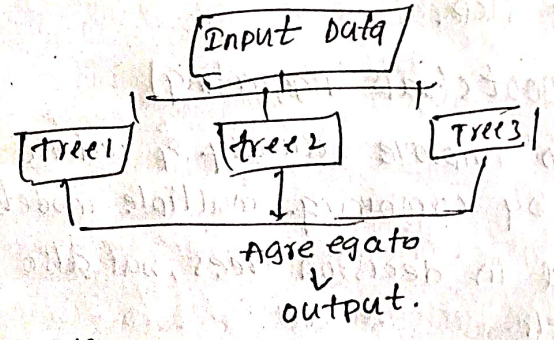
multiple trees:-
- work as a team (like a voting committee)
- Give more stable and accurate predictions.

Type of ensamble methods:-

1. Random forest

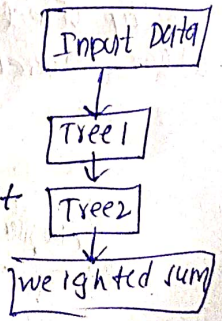
Builds many trees on random subsets of data and feature. final o/p is majority vote (classification) or average (regression)

- uses bagging + random feature selection
- Trees are trained independently
- works well for both classification & regression
- less overfitting than single tree.



2. Boosting:-
- Builds decision trees sequentially
- first tree is trained on full dataset

Boosting = many weak models → one strong model



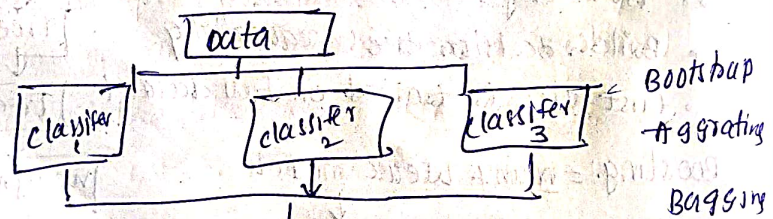
- Build models one after another (sequentially)
- first model is trained on normal training data
- next model focues on mistakes made by the previous model.
- Assign weights to data
  lower weight - correctly predicted
  Higher weight - misclassified
- New model focuces more on hard classes.
- This repeats
  - untill the model performs well
  - A set number of model is built
- Final prediction is a combination of all models (like weighted voting).

* To make a strong model by fixing errors step by step.

3. Bagging (Bootstrap Aggrating)

It used to improve accuracy & reduce overfitting by combining multiple models.

mostly used in decision trees, but also with other models.



How works:-
- create multiple datasets
- Train multiple models
- Train in parallel
- combine results.

Advantages :-
- Better accuracy, less overfitting
- more robust to noise

Disadvantages:-
- more computation, harder to interpret
- slower prediction.

Using multiple decision Trees
- Improves accuracy
- Reduces overfitting
- Gives stronger, more stable models.

Time Series :-

Time series analysis and forecast are crucial for predicting future trends, behaviors, and behaviours based on historical data.

A time series is a seaquence of data points collected, recorded or measured. at successive, evenly-spaced time intervals.

eg:- Intervals may be monthly, yearly, daily etc.

Time series forecasting is the process of using a statistical model to predict future values of a time series based on past results.

components of Time series data

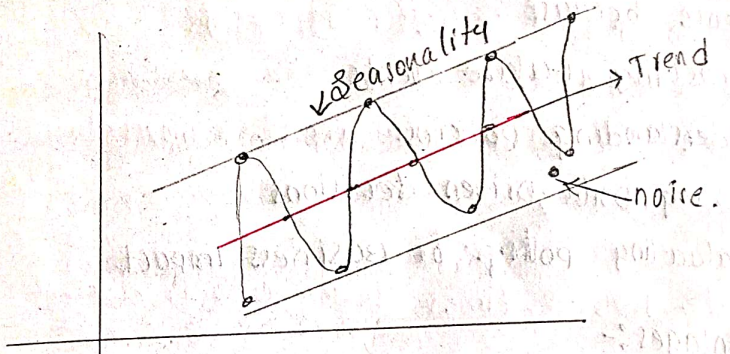
There are four main components.

1. Trend (T)
   A long-term upward or downward movement in the data.
   - shows overall direction of the data over time
   - can be increasing, decreasing or flat (no trend)
   e.g:- company revenue increasing steadily over 5 years.

2. Seasonality (S)
   A repeating pattern in data at regular intervals (days, months, quaters)
   - caused by seasonal effects, such as holidays, weather etc.
   - pattern is predicatble.
   e.g:- Ice cream sales increasing every summer.

3. Cyclical (C)
   similar to seasonality but not regular or fixed.
   - longer-term patterns often tied to economic cycles.
   - Not always predictable like seasonality.
   - usually spans years (e.g boom & recession)
   e.g:- Real estate prices over decades.

4. Irregular or Residual component (I)
   Represents random noise or fluctuation in the data
   - caused by unexpected events, like strike or natural disaster.
   - cannot be predicted.
   - e.g:- sudden drop in online traffic due to a server outage.



Characteristics :-
- Chronological order, - Sequential order
- Temporal components - Constant frequency
- Dynamic Nature.

Mathematical representation / Decomposition

A time series Y(t) can be expressed as.

1. Additive model

$$Y(t) = T(t) + S(t) + C(t) + I(t)$$

- used when the components are independent & linear
- Suitable when seasonal variation stays constant over time.

2. Multiplicative model

$Y(t) = T(t) \times S(t) \times C(t) \times I(t)$

- Used when components interact and amplify each other
- Suitable when seasonal variations increases/decreases with trend.

Importance of Time series Analysis:-

Time series analysis is crucial both business and science because

- forecasting future events
- Understanding patterns
- making Data-Driven decisions
- Evaluating policy or Business impact.

Advantages:-

1. Captures Temporal structure
2. Helps Discover Hidden insights
3. Enables Accurate forecasting
4. Useful in Automation.

~~ARIMA~~ Types of Time series methods

1. Decomposition methods like STL
2. Smoothing techniques
   moving average, exponential smoothing
3. Machine learning Models
   LSTSM, XGBoost.

Forecasting models of Time series:-
CAR, MA, ARIMA, ARIMA)

1. AR (AutoRegressive model) —(P)

uses past values to predict future

order of AR model is denoted by (P)

P= no. of lagged observations

AR(P) $y_t = C + \varphi y_{t-1} + \varphi y_{t-2} + \dots \varphi \varphi_{t-p} + \varepsilon$

2. MA (Moving Average) model —(q)

predicts current value using past forecast errors

q=no. of lagged error terms.

$y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} \dots \theta_q \varepsilon_{t-q}$

3. ARMA

AutoRegressive Moving Average

→ combines AR & MA model for forecasting

- captures both temporal dependences & error terms

→ order (P, q)

ARMA(P, q)

$y_t = \varphi_1 y_{t-1} + \varphi y_{t-2} + \varphi_p y_{t-p} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \dots \theta_q \varepsilon_{t-q}$

It only works if time series is stationary.

4. ARIMA

These is our main topic ARIMA

AR – Auto Regressive
I – Integrated
MA – moving Average

components:- $(\Delta y_t) = \varphi_1 (\Delta y_{t-1}) + \varphi_2 (\Delta y_{t-2}) \cdots + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots \theta_q \varepsilon_{t-q}$

1. AR – (P)
predicts future values using past values
P = no. of past values (lags) used.

2. I (Integrated) – d
- makes the data stationary (i.- removes trend)
uses differencing, like
$y_t - y_{t-1}$
- d = no. of times differencing is applied.

3. MA – (q)
uses past forecast errors to predict future values
q = no. of past errors used.

Notation ARIMA (P, d, q)

P – no. of AR terms
d – differencing degree
q – no. of MA terms.

Importance:- One of the most popular & relable models for time series forecasting
- Used in finance, sales, energy, weather predictions
- Helps make decision-driven approach based on trends.

When to use:-
use ARIMA when:-
- the data shows a trend
- The data is not stationary initially (becomes stationary after differencing)
- you want to forecast univariate time series.

Advantages:-
- Accurate for short term and medium-term forecasting
- Flexiable – fits many types of time series
- can handle data with trends & noise.

Measures of Forecast Accuracy

It tells us how close our predicted values (forecast) are to the actual observed values. They help us evaluate the performance of a time series model.

1. Mean Absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{t=1}^{n} |A_t - F_t|$$

A_t – actual value, F_t – forecast value
n – Total no. of observations.

- measures average absolute error b/w predicted and actual values
- simple and easy to interpret.

2. Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2$$

- squares the errors → larger error have more impact
- Helps detect larger deviations.

3. Root mean squared error (RMSE)

$$RMSE = \sqrt{MSE}$$

- Square root of MSE
- Penalizes big errors more than MAE

4. Mean Absolute percentage Error (MAPE)

$$MAPE = \frac{100\%}{n} \sum_{t=1}^{n} \left| \frac{A_t - f_t}{A_t} \right|$$

- Expresses error as a percentage
- Easy to understand "the forecast was off by 7% on average

limitation:-

Not reliable when actual values are very close to zero.

Why accuracy ~~matter~~ measure matters :-

- help to choose best model for forecasting
- Helps ~~to~~ detect overfitting or underfitting.

STL Approach :-

Seasonal-Trend decomposition using LOESS (Locally estimated scatterplot smoothing)

It is a powerful method to break down a time series into 3 clear parts:

1. Trend → long term direction of data
2. Seasonality - Regular, repeating patterns
3. Remainder (Residual) → leftover part (noise)

components :- If you have a time series Y_t, STL split it into

$$Y_t = T_t + S_t + R_t$$

- T_t : Trend component, S_t : seasonal component
  R_t : Remainder (random noise)

How it works :-

1. LOESS smoothing is used:
   LOESS is a flexiable smoothing technique that fits local regressions to small parts of data
   - It helps STL captase non-linear trends and complex seasonal effects.
2. STL works in iterations.
   First estimates trend → removes it → estimates seasonality → refines remainder.
   - This loop improves accuracy until results stabilize.

Advantages :-

- Handles changing seasonality
- Robust to outliers
- works with any season length.
- separates all 3 components

Real-life Applications :-

- Sales forecasting
- weather modeling
- Stock market trends
- weather modeling.

Extract Features from Generated model as Height, Average, Energy etc.

When we build a time series model (like STL, ARIMA), we can analyze the output components (trend, seasonal, residual) to extract useful numerical characteristics (called features).

These features help us:
- Compare different time series
- Detect patterns
- Improve forecasting
- Use them in machine learning models.

Common Features Extracted:-

Feature Name	Description
peak (Height)	– maximum value in the time series.
Through (Depth)	– minimum value in time series
Trend slope	– Rate of change of the trend component.
Average level	– mean value of the series or its trend.
Average energy	– sum of squares of the values used to capture intensity or power.
Seasonal strength	– How dominant the seasonal component is compared to the noise.
volatility	– standard deviation of the residuals
cycle length	– Average time between peaks or recurring patterns.
Auto correlation	– measures how correlated a series is with a lagged version of itself
Outlier count	– No. of unusual spikes or drops in the residuals.

Why this :-

- Height tells us about the peak demand/usage
- Energy tells how strong or intense seasonal effects are
- Volatility helps in risk assessement
- slope gives growth trend.

Used in:- . forecasting feature scales
. Identifying unusual behaviour.

Analyze for prediction

Time series involves understanding past data to accurately forecast feature values.
Once features like trend, seasonality, energy & volatility are extracted, we analyze there components to make informed predictions.

Steps to analyze :-

1. Decompose the Time series
Breake the series into components:
Trend, seasonality, Residuals.

2. Feature Analysis
Extract useful features
- Height, energy, volatility, Trend slope

3. Model fitting
fit a model to the time series:
- ARIMA / ARMA
- Exponential smoothing (ETS)

4. Evaluate Forecast Accuracy
use these metrics
MAE, MSE, RMSE etc.
lower error = better prediction

5. Generate forecast
- predict future time points using the fitted model
- Adjust based on seasonal & trend behaviour.

The better you analyze the components, the smarter your predictions. Always:
- visualize components
- check feature importance
- choose model that best match your dat behaviour.

Advantages:-
- Improved Accuracy
- Simplifies complex Data.
- faster model training
- Model-Agnostic
- Robust to Noise

Applications:-
- Finance
- Healthcare
- Energy
- Retail
- manufacturing.

(or)
once features are extracted, we use them to
- Train machine learning model
- Cluster Time series
- Detect Anomalies.

unit -V

Data Visualization

- Pixel-oriented visualization Techniques
- Geometric projection visualization Techniques
- Icon Based visualization Techniques
- Hierarchical visualization Techniques
- visualizing complex Data & Relations.

Data visualization

It is the graphical representation of information and data. By using visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data.

Why is it important ?

- Human process visuals faster than raw data
- Helps in identifying, trends, patterns, & outliers quickly
- Enables better decision-making through insights.
- Makes complex data easier to understand.
- useful in storytelling with data, especially in business and analytics.

Objectives :-

- make data clear and intuitive
- show relationships b/w variables
- support data driven decision making
- Detect anomalies & trends.

common Tools :-

- Tableau
- power BI
- Excel
- python
- R
- D3.js.

Advantages :-

- speeds up data analysis
- makes reports engaging & easy to interpret
- Useful for both technical and non-technical users
- helps in real time monitoring through dashboards
- Improves communication across teams & stakeholders.

-Applications :- Business Intelligence, Healthcare, -finance, marketing, Government.

categorization of visualization methods

- Pixel oriented
- Geometric projection
- visualizing complex data and relations.
- Icon based
- Hierarchical

Pixel-Oriented Visualization Techniques

Pixel-oriented visualization is a method that uses color-coded pixels to represent individual data values from large and high dimensional datasets.

It ~~deals~~ is ideal when you want to see many dimensions and records on a single screen.

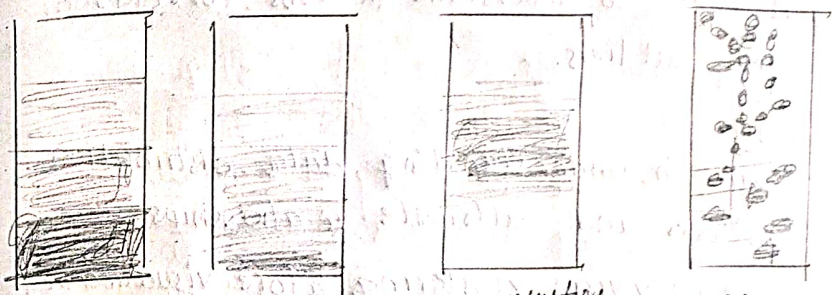
→ It is a way to show large amounts of data using tiny colored squares (pixels)

each pixel represents one data value and the color of the pixel tells you how high or low that value is.

How it works :-

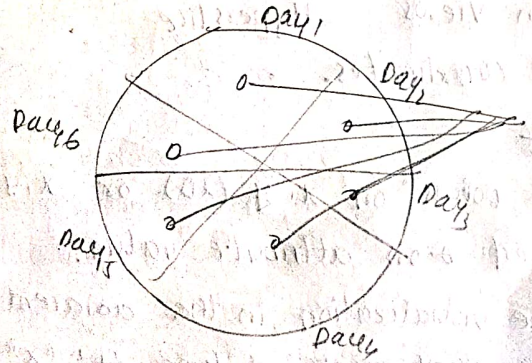
- each datavalue is shown as a colored pixel
- for a dataset with m dimensions, create m windows (one per dimension)
- A record's m value are shown as m colored pixels, one in each window.
- All windows, follow the same global order usually by sorting on one key attribute
- this method allows large datasets to be shown compactly on screen.

e.g :
- Customer data 4 dimensions: income, transaction volume, age et
- data is sorted by income
- visualization show how other attributes relate to income
- Helps in spotting patterns & correlations.



Laying out pixels in circle segments.

To save space and show the connection among multiple dimensions, space filling is often done in circle segement



Advantages
- compact display
- Detailed view
- multiple dimensions
- Fast comparision

Drawbacks
- color interpretation
- Requires sorting
- overload risk
- limited interaction.

Geometric projection visualization techniques.

To visualize multidimensional data by protecting it into 2D or 3D space, making it easier to understand patterns, correlations and outliers.

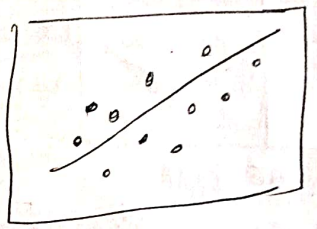
Key:-
- Helps in understanding data distribution, outliers and attribute relationships.
- Often requires clustering before visualization to reduce cluster in large datasets.

Methods:-
- Direct visualization, - scatterplot and scatterplot matrices
- Landscapes - projection pursuit technique.
- prosection views - Hyperslice
- parallel coordinates.

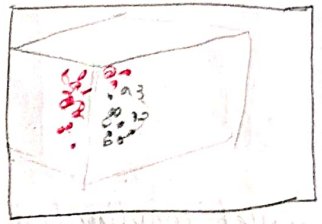
Scatterplot
- plots data points on x, y (2D) or x, y, z (3D) axes based on attribute value.
- Through the visualization, in the adjacent figure we can see that points of types '+' "x" tend to be collected.

use:- Best for showing correlation blw 2 or 3 attributes

limitation:- Get clustered with large or high dimensional data.



2D scatter plot

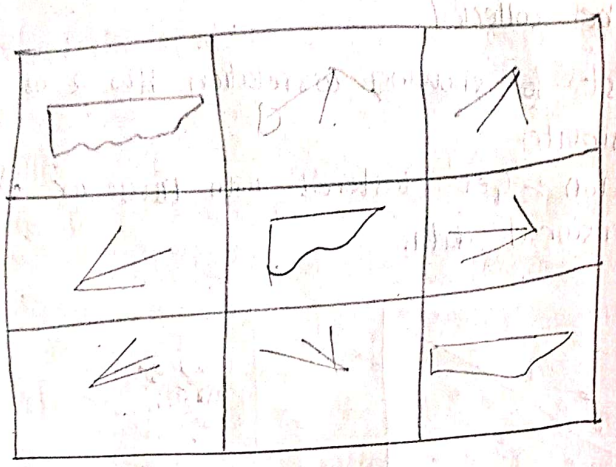


3D scatter plot.

Scatterplot matrix (SPLOM)
- A scatterplot matrix is a grid of scatter plots used to used to visualize relationships blw multiple attributes (dimensions) of a dataset.
- for a dataset with k attributes, SPLOM shows k x k 2D scatter plots
- each cell shows a plot of one attribute vs another.
- Diagonal cells (eg A vs A, B vs B) are usually,

Hyperslice
- shows 2D slices of a multidimensional function as a matrix
- each slice shows how the function behaves with 2 variables while others are fixed.
- Advantage: easy interaction; users can explore dimensions directly.

use:- Great for scaler fn & mathematical data analaysis.

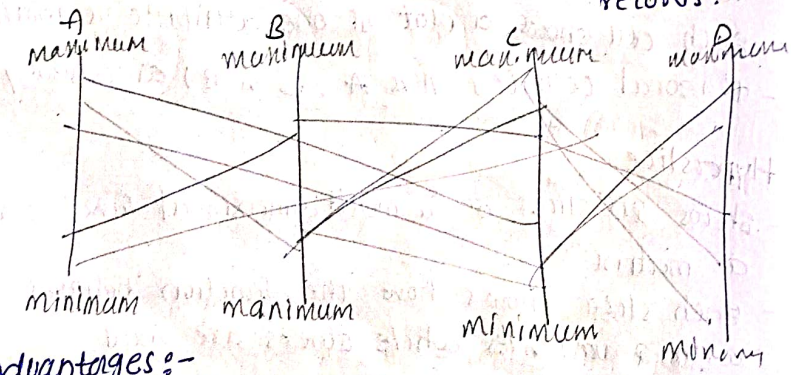


parallel coordinates

How it works: Draws parallel axes for each dimensions

- Each data point is a line crossing the axes at its attribute values.
- Use : Useful for identifying patterns and clusters across many dimensions.

limitations:- Becomes hard to read with too many records.



Advantages:-

- Reveals hidden patterns, clusters and correlations
- suitable for multivariant data
- Interactive exploration.

Drawback:-

- cluster with large dataset
- can be hard to interpret without preprocess & filtering
- limited scalability.

Applications:-

- outlier Detection
- correlation analysis
- cluster Identification
- Dimensionality reduction
- finance and Business
- Heath care & Bioinformatics
- Engineering & manufacturing
- Social Network analysis.

Icon-Based visualization techniques.

These techniques use small icons to represent multidimensional data. Each icon visually encodes several variables using shapes, angles, sizes or colors.

- Instead of showing data in traditional graphs or charts, each data record is converted into a small icon
- Each part of the icon reflects a different data value
- Icons are arranged in a grid or layout for comparision.

Common Techniques are:-

* Chernoff faces
* Stick figures
* Star Glyphs
* Color Icons.

General techniques :- shape coding, color icons, Title bars.

Chernoff Faces

Each data attribute controls a part of cartoon face (e.g. mouth, face, eyes $. Human eyes easily detect patterns or anomalies in facial expressions.

- use cartoon faces to represent data
- each facial feature (like eyes, mouth, nose, face shape) represents a variable.
- can display up to 18 dimensions
- easy for humans to notice patterns or anomalies using facial expression differences.

Ex- Each face = one record, eyesize → age, mouth curve → income etc.

- viewing large data can be tedious
- Chernoff faces make the data easier for users to digest.

Stick figure

- Represent data as a 5-part stick figure (4 limbs + body)
- 2 variables are mapped to x an y axes (positions)
- Other dimensions are mapped to the lengths or angles of the arms and legs.
- Good for visualizing motion or variation b/w records.

Advantages :-

- multidimensional display
- Human pattern recognization
- compact representation
- Intuitive Interpretation.

Limitations / Drawbacks

- limited precision
- difficult scaling
- visual cluster
- Interpretation bias

Applications:-

- pattern discovery & customer profiling
- medical data analysis
- psychology & social science
- Education & research.

Hierarchical visualization Techniques

Hierarchical visualization is used to display data that is organized in a tree-like structure (parent-child relationships). These techniques help us easily to understand nested relationships like :-

- Organization charts
- file / foleder systems
- Customer segmentation
- clustered data.

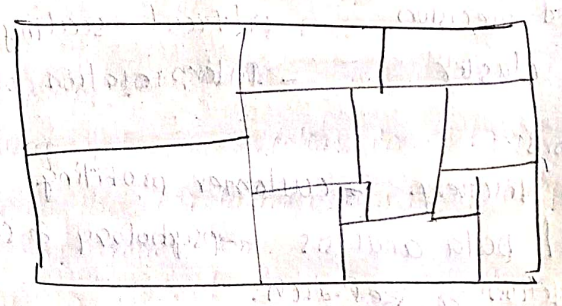
common Techniques:-

1. Tree Map:-

A rectangular layout where each rectangle (box) represents a data point or category

working:-
- Nested boxes shows hierarchy
- size and color of the boxes represent data values (like cells, size, etc)

use case:-
visualizing resource usage, financial data, or folder structure.

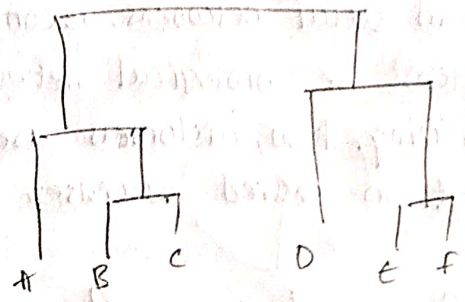


2. Dendrograms:-

A tree like diagram that shows how clusters are formed.
used in Hierarchical clustering.

working:-
- starts with individual data points
- merges them step by step into clusters based on similarity
- Continues until one large cluster remains.

Two main clustering approaches:
- Agglomerative (bottom-up):- starts with single data points and merge clusters
- Divisive (top-down):- start with all data in one cluster and split into smaller ones



Applications:-
- customer segmentation
- organizational charts
- Gene expression data (bioinfosmatics)
- Document clustering.

advantages
- shows nested structure & relationships clearly
- Helps with cluster analysis

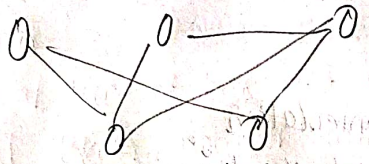
limitations:-
- may become cluttered with too many levels
- Not ideal for flat or non-hierarchial data
- Requires proper scaling for readability.

Visualizing Complex Data and Relationships:

when dealing with complex data - such as large, multi-dimensional, or interconnected datasets. we use advanced visualization techniques to reveal hidden patterns, relationships and insights.

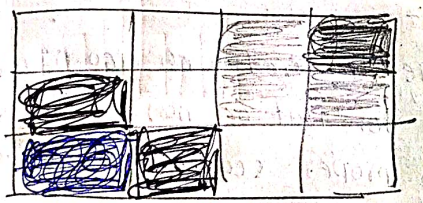
Techniques to visualize complex data

1. Network Graphs
   - show relationships between items (nodes & edges)
   - common in social networks, recommendation systems, or biological networks.
   e.g:- Visualizing how customers are connected based on shared purchases.



random

2. Heat maps:-
   - Display values as colors in a grid
   - used for correlation matrices, activity logs, gene generation.
   - Easy to detect clusters, anomalies or high/low values

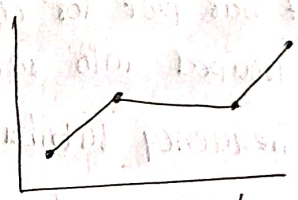


colored heat map.

3. Dimensionality Reduction techniques.
   used To reduce complex high-dimensional data into 2D or 3D for visualization:
   - PCA
   - t-SNE (t-Distributed Stochastic Neighbour Embedding)
   - UMAP.

4. 3D scatter plots & Interactive Dashboard
   - used to explore multiple variables with zooming, rotation & filtering
   Tools like power BI, Tableau and ploty.

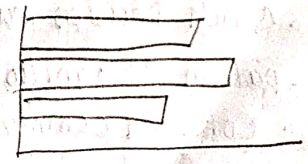
5. line plot
   A line plot shows data points connected by straight lines
   used for showing trends over time
   e.g:- Monthly revenue for 1 year.



line plot

6. Bar plot / Bar Graph:-
   Represents categorical data with rectangular bars.
   The height or length of the bar shows the value.
   Types:- vertical bars : Bar plot
   Horizontal : Bar graph
   used for : comparing categories.

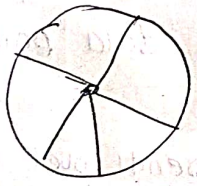
Bar plot



Bar graph

7. pie chart

A circular chart divided into slices to show proportions. used for displaying part to whole relationship showing percentages or ratios.



8. Histogram

A special type of bar plot for continous numerical data, grouped into ranges (bins)

Used for : showing frequency distributions

Best for : Identifying data spread, skewness, outliers & normality

Advantages :-

- Improved understanding
- Quick Decision making
- pattern Recognition
- Enhanced communication
- Data exploration
- Time saving.
- Error Detection.

ETL (Extract, Transform, Load)

3 step process used in data warehousing and analytics to move and prepare data for analysis

1. extract
   collect data from different sources

2. Transform
   - clean & convert the data into a standard usable format (removing duplicates, formats)

3. Load
   store the transformed data into a target database or data warehouse for reporting and analysis

popular ETL tools

- Informatica powercenter
- IBM Infosphere Data stage
- Apache spark
- pentaho (kettle).

Word Cloud :-

A word cloud is a visual display of words from document or dataset.

- Bigger words appear more frequently in text
- Helps highlight keywords, themes, trends in textual data
- common in text mining & social media analysis.