

PART - A

1.a) Mention any two issues affecting in data Quality.

* Missing data :- Missing data can lead to biased analysis, reduced statistical power and potentially incorrect conclusions.

* Inconsistent Data :- Issues like formatting issues, duplicate errors can lead to errors in analysis.

b) What is the role of Sensors in data collection?

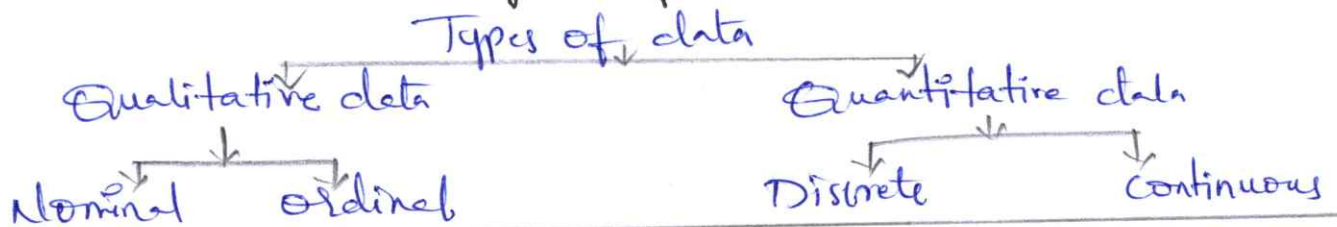
Sensors enabling real-time monitoring like sensor facilitate continuous, real-time data collection for decision-making.

* Capturing physical parameters, Improving accuracy.

c) State any one application of analytics in business?

Analytics in business is customer Segmentation and Targeting. By analyzing customer data such as purchase, history, and behavior, business can identify distinct customer segment.

d) What are the main types of data?



e) What is Variable rationalization?

Variable rationalization is the process of selecting and reducing the number of variables in a dataset to improve model performance. mainly include PCA and feature selection methods.

f) What is Least Square Estimation?

LSE is a statistical method used to estimate the parameters of a linear regression model by minimizing the sum of squared errors between observed and predicted values.

g) Define ETL approach.

ETL [Extract, Transform, Load] is a data integration approach used to extract data from multiple sources, transform it into a standardized format, and load it into a target system.

h) What is ARIMA model?

ARIMA [Auto Regressive Integrated Moving Average].

It is a statistical model used for forecasting and analyzing time series data.

i) Define complex data relations.

Complex data relations refer to intricate connections between variables, involving non-linear, hierarchical requiring advanced analytical techniques to identify and model.

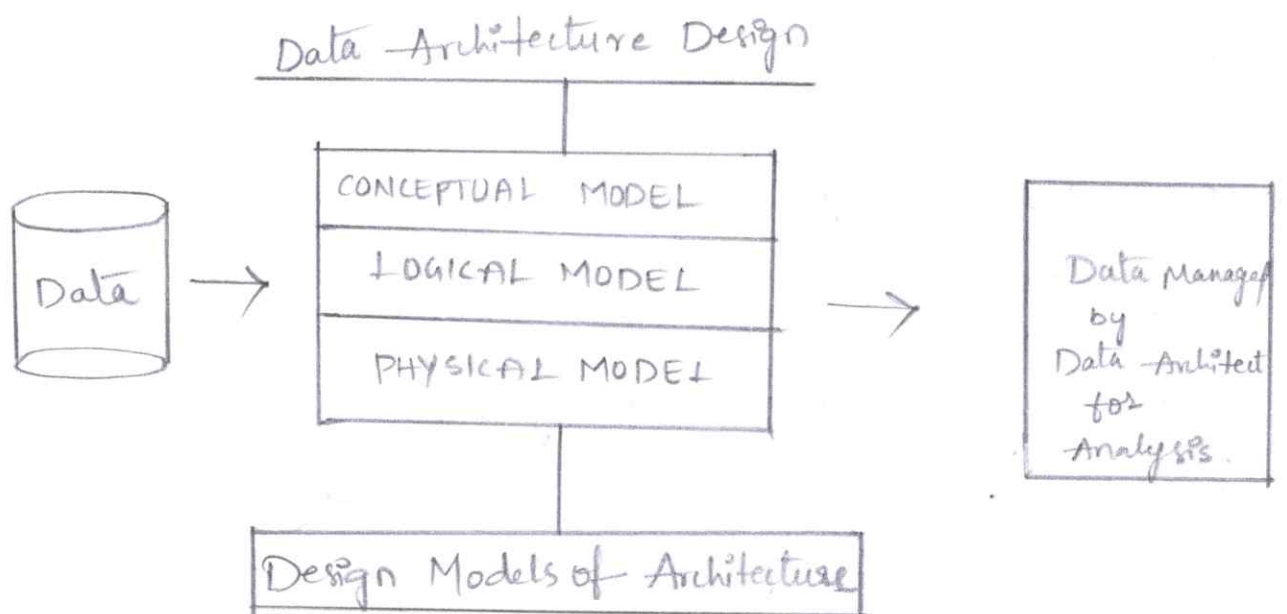
j) Define pixel-oriented visualization.

Pixel-oriented visualization is a technique where each data value is mapped to a pixel on the screen, allowing for the visualization of large datasets by representing individual data points as colored pixels.

PART-B

Q. Discuss the components of data architecture in detail and explain how data is managed for analysis.

- * Data architecture is set of standards which are composed of certain policies rules, models and standards.
- * And which manages, what type of data is collected, from where it is collected and the arrangement of collected data, and securing the data into the systems for further analysis.
- * Data architecture describe the type of data structures applied to manage data and it provide an easy way for data preprocessing.
- * Data architecture is formed by dividing into three essential models and then are combined.



Key components of data Architecture :-

* Data architecture is a critical component of data analytics, and it involves designing, implementing, and managing the structure and flow of data within an organization.

* Data Sources :-

* Internal data Sources :- Data generated within the organization such as transaction data, customer data.

* External data Sources :- Data obtained from external sources, such as social media, market research and public data.

* Data Ingestion (or) collection / Integration :-

* Data collection :- Gathering data from various sources, including batch processing.

* Data Integration :- combining data from multiple sources into a unified view.

* Data Storage :-

* Data warehouses :- centralized repositories for storing and managing data.

* Relational databases :- Traditional databases designed for structured data.

* Data Processing :-

* ETL (Extract, Transform, Load) :- Extracting data from sources, transforming it into a usable format.

Data Management for Analysis :-

* Data Management is the process of managing the tasks like extracting data, storing data, transferring data, processing data and then securing data with low-cost consumption.

* Main motive of data management is to manage and safeguard the people's and organization data in an optimal way. So that they can easily create, access, delete, and update the data.

3.a) Discuss various data quality issues in missing values.

* Missing values are a common data quality issue that can significantly impact the accuracy and reliability of data analysis.

* Missing completely at Random (MCAR).

* Missing at Random (MAR)

* Missing Not at Random (MNAR)

* Missing values like NA (not available) and NULL and NaN.

* NA (not Available): A missing value that is not available (or) not applicable.

* NULL: A missing value that is Unknown (or) Undefined.

* NaN (not a Number): A missing value that is not a number.

Ex - division by zero.

MCAR [Missing completely at Random].

* Missing values are randomly distributed throughout the dataset, and there is no pattern (or) relationship between missing values and other variables.

MAR [Missing at Random].

* Missing values are related to observed variables, but not to the missing variable itself.

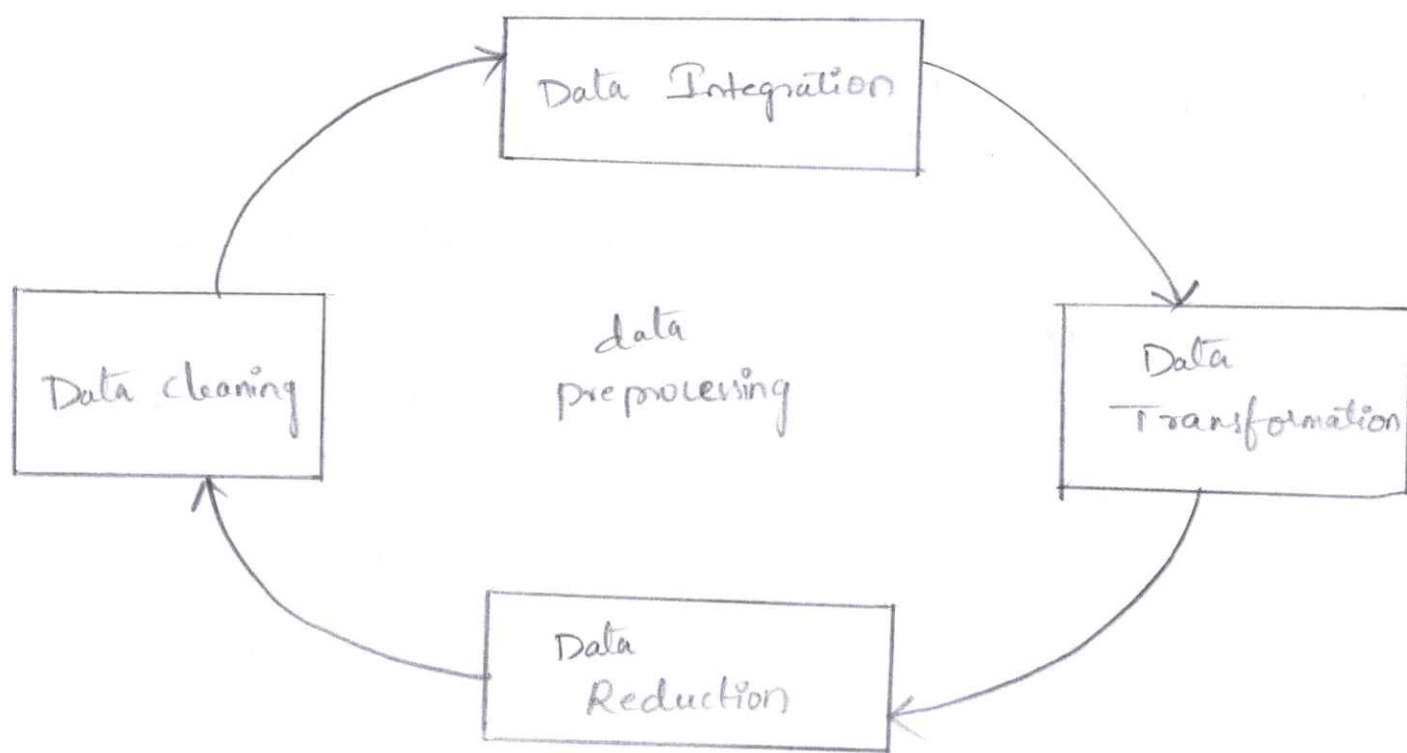
MNAR [Missing Not at Random].

* Missing values are related to the missing variable itself, and there is a systematic reason for the missing data.

3.b) What is data preprocessing? Describe its stages?

* The process of transforming raw data into an understandable format.

* In this major tasks 4.



Data cleaning:- process of removal of incorrect, incomplete, inaccurate data also replaces missing data.

* Handling missing values.

* Handling noisy data:

Data Integration:- Multiple heterogeneous sources of data are combined into single dataset.

* 2 types of Integrations are 1. Tight coupling.

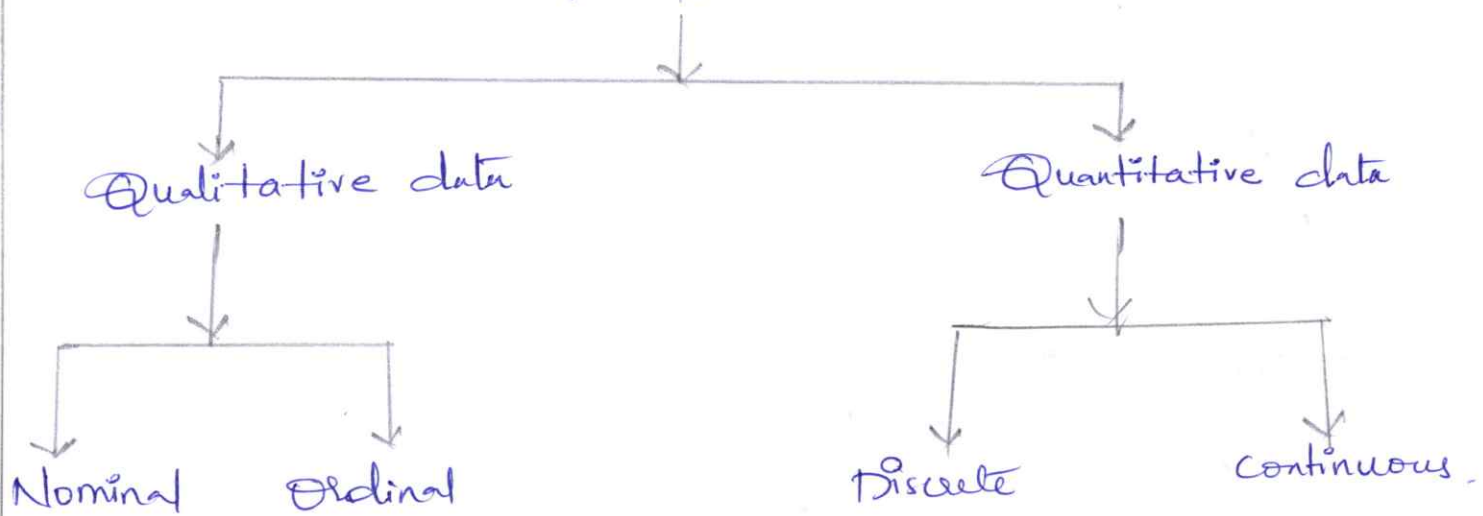
2. loose coupling.

Data Reduction:- Volume of data is reduced to make analysis easier.

Data Transformation:- Data is transformed into appropriate form suitable for mining process.

4.a) What are the different types of data and variables?

Types of data



Qualitative data :-

* This type of data can't be counted (or) measured easily using numbers and therefore divided into categories.

Nominal :- Set of values that don't possess a natural ordering. Ex - color of a smartphone, and gender of a person.

Ordinal :- values have a natural ordering while maintaining their class of values.
Ex - clothing brand. Size [small < medium < large].

Quantitative data :-

* This type of data tries to quantify things and it does by considering numerical values that make it countable in nature.

Ex - price of smartphone.

Discrete :- numerical values which fall under are integers number are placed this category.

Continuous :- fractional numbers are considered as continuous values. Ex - wifi - frequency.

Types of variables :-

- * 4 types of variables
- * Categorical
- * Date Time
- * Mixed
- * Numerical

Date Time Variable :- deals with the date & time aspects.

Ex - Date of Application.

Mixed variable :- collection of multiple values for multiple observations.

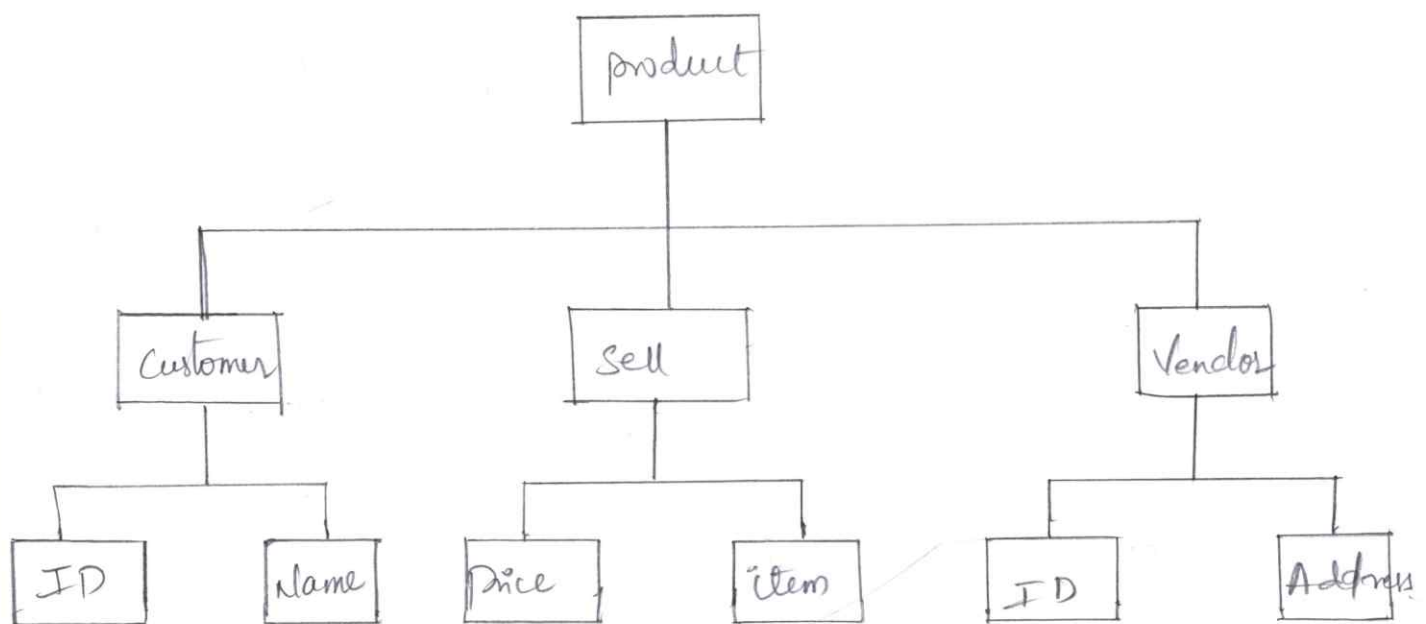
Ex - performance of student.

4b) Explain in detail about data Modelling Techniques.

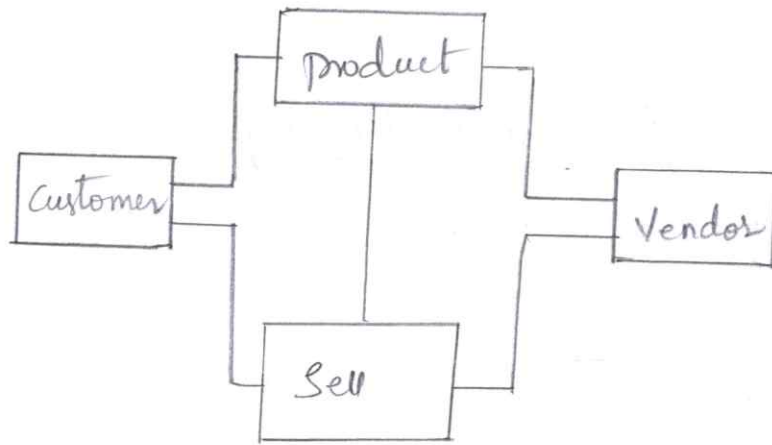
* There are various techniques to achieve data modeling successfully.

* Some popular data modeling techniques include Hierarchical, Relational, Network, Entity-relationship and Object-oriented.

Hierarchical Data Modeling Technique



Network data Modeling Technique.



- * Network data Model technique is a flexible way to represent objects and Underlying relationships between entities.
- * Here objects are represented inside nodes and the relation-ship b/w the nodes is illustrated as an edge.

5) Explain the process of data analytics and discuss its applications in business decision-making.

- * The process of data Analytics involves several steps that help organizations extract insights from data and make informed decisions.

- * **Problem Identification:** Identify the business problem (or) opportunity that data analytics can help address.

- * **Data collection:** Gather relevant data from various sources, including databases, files, and external sources.

- * **Data preprocessing:** Clean, transform, and prepare the data for analysis.

- * **Data Analysis:** Apply statistical and analytical techniques to identify patterns, trends and insights.

- * **Decision making:** Use the insights to inform business decisions.

* Deployment: Implement the decisions and monitor the outcomes.

Applications in business decision-making:-

* Data Analytics has numerous applications in business decision making include

* Market Segmentation:- Identifying target customer segments based on demographic, behavioral, and transactional data.

* customer churn prediction:- predicting which customers are likely to churn based on historical data and behavioral patterns.

* Sales forecasting: predicting future sales based on historical data, seasonal trends, and market conditions.

* Supply chain optimization: optimizing supply chain operations by analysing data on inventory levels, shipping routes and supplier performance.

* financial analysis: Analysing financial data to identify trends, risks and opportunities for cost savings.

* Risk Management: Identifying and mitigating risks by analysing data on market trends, customer behaviour and operational performance.

- 6) Apply a regression model to forecast future financial performance using a real-time dataset of company revenue and marketing expenditure. Analyze the regression results and interpret how marketing investment impact revenue growth.

Marketing Spend (₹ Lakhs)		Revenue (₹ Crores)
Year		
2020	30	5.5
2021	35	6.1
2022	38	6.8
2023	40	7.2
2024	45	8.0

Step 1 :- Define the problem and data.

* We need to apply a regression model to forecast future financial performance based on the given dataset of company revenue and marketing expenditure.

Step 2 :- Choose a regression model.

* We will use a simple linear regression model to analyze the relationship between marketing spend and revenue.

$$* \text{Revenue} = \beta_0 + \beta_1 \times \text{marketing spend} + \epsilon.$$

Step 3:- Calculate Regression coefficients.

$$\bar{x} = \frac{30+35+38+40+45}{5} = 37.6$$

$$\bar{y} = \frac{5.5+6.1+6.8+7.2+8.0}{5} = 6.72$$

linear regression $B_0 = \bar{y} - B_1 \bar{x}$, $\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$

Step 4:- compute Deviations and Sums.

$$B_1: (30-37.6)(5.5-6.72) + (35-37.6)(6.1-6.72)$$

$$= (-7.6)(-1.22) + (-2.6)(-0.62) + (0.4)(0.08)$$

$$= 9.272 + 1.612 + 0.032 + 1.152 + 9.472 = 21.5$$

$$\sum (x_i - \bar{x})^2 = (-7.6)^2 + (-2.6)^2 + (0.4)^2 + (2.3)^2$$

$$= 57.76 + 6.76 + 0.16 + 5.29 + 54.76 = 125.2$$

Step 5:- Calculate coefficients.

$$B_1 = \frac{21.54}{125.2} \approx 0.172$$

$$B_0 = 6.72 - 0.172 \times 37.6 \approx 6.72 - 6.467 = 0.253$$

Step 6 + Interpret Results.

$$\text{Revenue} = 0.253 + 0.172 \times \text{marketing spend.}$$

* Indicates for every ₹ 1 Lakh increase in marketing spend, revenue increases by approximately ₹ 0.172 Crores.

Step 7 + forecast future performance.

Ex ₹ 50 Lakhs:

$$\text{Revenue} = 0.253 + 0.172 \times (\text{times } 50) = 0.253 + 8.6 = 8.853 \text{ Crores.}$$

7.

Explain about regression types and detailed note on model construction in logistic regression. (7)

Regression :-

* The evaluation of relationship between two (or) more variables is called Regression Analysis. It is a statistical technique.

* It generates a regression equation where the coefficient correspond to the relationship b/w each independent and dependent variable.

Types of Regression :-

- Linear Regression
- Logistic Regression
- polynomial "
- Support vector "
- Decision Tree "
- Random Forest "
- Ridge "
- Lasso "

Logistic Regression Model construction :-

* Logistic regression is a widely used statistical technique for binary classification problems.

- * Data preparation
- * Model specification
- * Model Estimation.

→ Model Evaluation

→ Model Validation.

* Logistic Regression is another Supervised Learning algorithm which is used to solve the classification problems.

* In classification problems, we have dependent variables in a binary (or) discrete format such as 0 (or) 1

* It mainly works with the categorical variable such as 0 (or) 1, Yes (or) No, True (or) False.

* Logistic regression uses Sigmoid function.

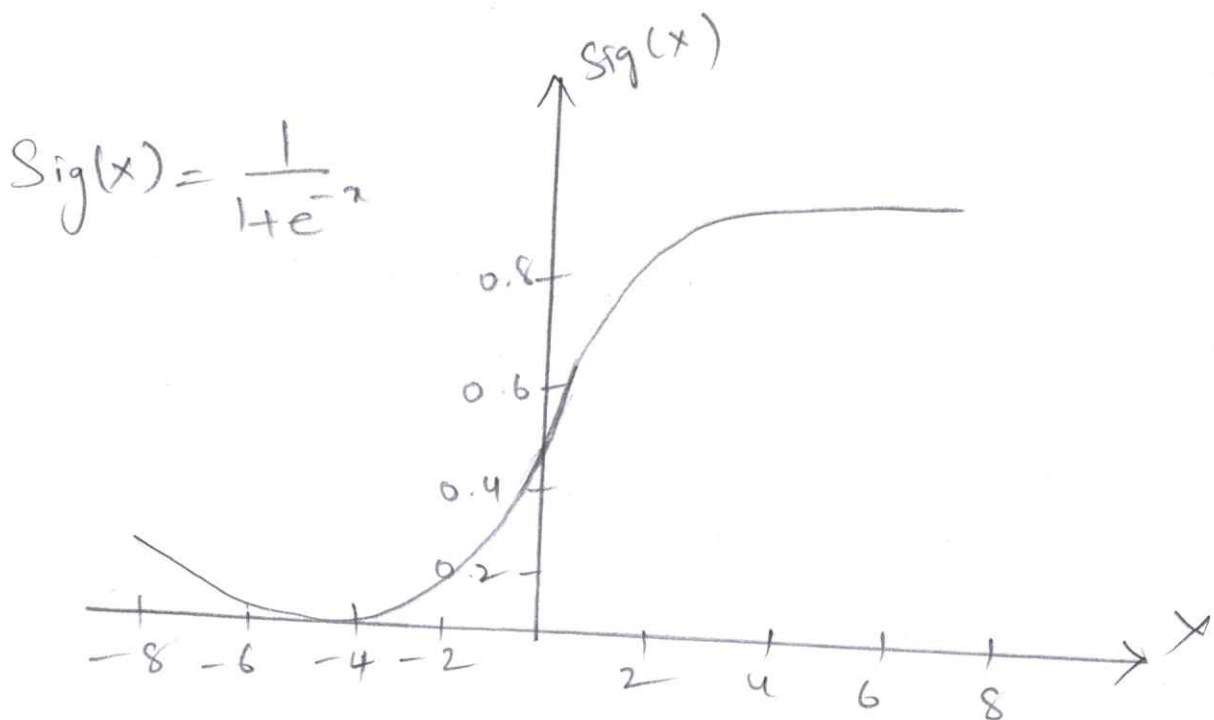
which is complex cost function.

* represented as
$$f(x) = \frac{1}{1 + e^{-x}}$$

* $f(x)$ = o/p between the 0 and 1 value.

* x = i/p to the function

* e = base of natural logarithm.



8) Explain time series analysis methods and measures of forecast accuracy.

* Time series is a set of observations which are taken at a particular time, at a regular interval.

ARIMA Method:-

* Auto regressive Integrated moving Average

* AR model are the models in which the value of a variable in current period is related to its value in the previous period.

* MA model: Moving Average models are the models in which the value of a variable in current period is related to value of the residual in the previous period.

* ARMA Model:- Auto Regressive moving Average model are the model in which the value of a variable in current period is related to it's own value in the previous period as well as values of the residual in the previous period i.e., ARMA is a combination of both autoregressive terms (p) and moving Average (q) terms.

* ARIMA model:- Auto Regressive Integrated moving Average. It is a popular model for forecasting time series data.

* SARIMA model:- seasonal ARIMA it is an extension of ARIMA for seasonal data.

* Time series components: Trend, seasonality, cyclical, Irregular.

Exponential Smoothing :- A family of models that use weighted averages to forecast data.

Measures of forecast accuracy :-

* Measures of forecast accuracy are used to assess the performance and reliability of forecasting models.

* Here are some commonly used measures of forecast accuracy.

* MAE [Mean Absolute Error] :-

MAE calculates the average absolute difference b/w the forecasted values and the actual values. It measures the average magnitude of errors without considering their direction.

* formula of MAE = $(1/n) * \sum |Actual - Forecast|$

* MSE [Mean Squared Error] :-

MSE calculates the average of the squared difference between the forecasted values and the actual values. It [measures the average of] penalizes larger errors more than the MAE and is commonly used in many forecasting applications.

* formula of MSE = $(1/n) * \sum (Actual - Forecast)^2$

* RMSE [Root Mean Squared Error] :-

RMSE calculates the square root of MSE and provides an interpretable measure in the same units as the original data. It is widely used and is helpful in understanding the average size of errors.

* formula of RMSE = $\sqrt{MSE}$.

9) Write short note on Regression vs Segmentation and STL approach in forecasting.

Regression :-

* Regression analysis is a statistical technique used to model the relationship between a dependent variable and one (or) more independent variables.

* The goal of regression analysis is to identify the nature and strength of the relationship between the variables. and to make predictions about the dependent variable based on the values of the independent variables.

Segmentation :-

* Segmentation involves dividing a larger dataset into smaller groups (or) segments based on similarities (or) differences in the data.

* The goal of segmentation is to identify patterns and relationships within the data that may not be apparent when analyzing the dataset as a whole.

* both are different techniques in data Analytics.

STL approach :-

- * STL (Seasonal and Trend decomposition Using Loess).
- * This approach is a popular time series decomposition method that separates a time series into three components: trend, seasonality and residual.
- * It is particularly useful when dealing with the time series that exhibit both trend and seasonal patterns.

STL approaches :-

* Seasonal component :-

- * The STL approach first estimates the seasonal components of the time series.
- * It does this by applying a locally weighted regression (loess) to smooth out the series and identifying seasonal patterns.

* Trend component :-

- * After extracting the seasonal component, the STL approach estimates the trend component.
- * This is done by applying another loess smoothing.

* Residual components :-

- * Residual components represents the remaining variation in the time series after removing the seasonal and trend components.

10) Analyze hierarchical visualization techniques and its usage in Healthcare.

* Hierarchical visualization techniques are used to represent complex data in a hierarchical structure making it easier to understand and analyze.

* Here are some techniques to represent hierarchical data.

* Tree maps :-

A tree map is a rectangular visualization technique that displays hierarchical data as a set of nested rectangles.

* Sunburst charts :- A Sunburst chart is a circular visualization technique that displays hierarchical data as a set of concentric rings.

* Network diagrams :- A Network diagram is a visualization technique that displays hierarchical data as a set of connected nodes and edges.

* Tree diagrams :- Tree diagrams also known as dendrograms or hierarchy diagrams.

* Use like a tree-structure represent hierarchical relationships.

* data as a set of connected nodes.

Usage in Healthcare :-

* Hierarchical visualization techniques are widely used in healthcare to analyze and visualize complex data. Such as -

* patient data :- Hierarchical visualization techniques can be used to analyze patient data, such as medical history, diagnoses, and treatments.

* Disease outbreaks :-

Hierarchical visualization techniques can be used to track disease outbreaks and identify patterns and trends.

* Healthcare Networks :-

Hierarchical visualization techniques can be used to analyze healthcare networks, including patient flow, resource allocation, and supply chain management.

* Clinical trials :-

Hierarchical visualization techniques can be used to analyze clinical trial data, including patient outcomes, treatment efficacy, and adverse events.

11) Discuss in detail about geometric and icon-based visualization methods. (11)

Geometric visualization Methods :-

* Geometric visualization methods are used to represent complex data in a geometric format, making it easier to understand and analyze.

* Some common geometric visualization methods.

* mainly it transform three dimensional (3D) geometric objects onto a two-dimensional (2D) plane.

* Perspective Projection :-

* It mimics how the human eye perceives objects in 3D space.

* It creates the illusion of depth by projecting center.

* This technique is used in architectural renderings, computer graphics.

* Orthographic Projection :-

* It is a type of parallel projection that preserves the relative sizes and shapes of objects.

* It projects each point of a 3D object onto 2D plane.

* Axonometric projection :-

* It is a family of parallel projection that maintain relative sizes and shape of objects.

* and also main equal scaling along each axis.

* Oblique projection :-

* It is also type of parallel projection that combine orthographic projection with a skew transformation.

Icon-based Visualization method :-

* Icon-based visualization methods use icons (or) symbols to represent data, making it easier to understand and analyze complex information.

* These techniques are effective for conveying information quickly and intuitively.

* Icon Arrays :-

* In this use individual icons to represent discrete values.

* Icons are arranged in a grid (or) matrix, with each icon representing a specific quantity.

* Bar Icons :-

* Bar Icons are similar to bar charts but use graphical icons instead of traditional bars.

* Glyphs :-

* Glyphs are simple, abstract icons (or) symbols that represent specific concepts (or) data points.

* Chernoff faces :-

* A way to display on a two-dimensional surface
Ex let x be eyebrow slant, y be eye size, z be nose length, etc.

