Section A

1. **Explain the following Questions (15 marks total):**
    - (a) Random variable Genetic Algorithm (6 x 2.5 = 15)
    - (b) Goodness of fit - likelihood ratio test
    - (c) Mean stationarity
    - (d) Gaussian mixture model
    - (e) Sizing sketches
2. **What is Data pre-processing? What is exploratory data analysis? Explain its types and how does it work?** (15 marks)
3. (a) **Differentiate between covariance and correlation with example.** (8 marks)

   (b) **Explain in detail canonical correlation.** (7 marks)

Section B

4. **What is Hypothesis Testing? How is Hypothesis Testing Used in Data Science? Where and When to Use Hypothesis Test?** (15 marks)

5. Answer the following: (15)

(a) Is PCA or SVD better for dimensionality reduction?

b) What is the difference between t-SNE and PCA for dimensionality reduction?

Section-C

6. What is the AR model in time series? How do we interpret Autoregressive Model? Explain its Process. (15)

7. Describe the Autoregressive Integrated Moving Average (ARIMA) model. What are the three main components of the ARIMA model, and how are they combined to model a time series? (15)

Section-D

8. Compare the efficiency and accuracy of bootstrapping and Monte Carlo methods in estimating statistical parameters or solving numerical problems. Under what circumstances would you choose one method over the other, and why? (15)

9 Explain the following: (15)

(a) EHR data

(b) Price optimisation in retail

(c) Demand forecasting

(a) Random Variable Genetic Algorithm

A **Genetic Algorithm (GA)** is an optimization technique inspired by natural selection. It evolves a population of candidate solutions through selection, crossover, and mutation to find the best solution.

(b) Goodness of Fit - Likelihood Ratio Test

The **Likelihood Ratio Test (LRT)** compares the goodness of fit between two models by evaluating the ratio of their likelihoods. A significant difference suggests that the more complex model provides a better fit to the data.

(c) Mean Stationarity

Mean stationarity refers to a time series where the mean remains constant over time, and the autocovariance depends only on the lag, not on time.

(d) Gaussian Mixture Model

A **Gaussian Mixture Model (GMM)** is a probabilistic model assuming that data is generated from a mixture of several Gaussian distributions. It uses the Expectation-Maximization algorithm to estimate parameters.

(e) Sizing Sketches

Sizing sketches are preliminary drawings used in manufacturing and engineering to determine the appropriate dimensions and specifications of components or systems.

1. **What is Data pre-processing? What is exploratory data analysis? Explain its types and how does it work?** (15 marks)

Ans **Data Preprocessing**

Data Preprocessing is the process of cleaning, transforming, and organizing raw data into a structured format suitable for analysis, machine learning (ML), and artificial intelligence (AI) applications. It aims to enhance data quality, improve model performance, and reduce computational complexity.

Key Steps:

1. **Data Cleaning:** Handling missing values, removing duplicates, and correcting errors.
2. **Data Transformation:** Standardizing, normalizing, or encoding categorical data.
3. **Data Integration:** Merging data from multiple sources into a unified dataset.
4. **Feature Scaling:** Adjusting numerical values to promote fair weightage in ML models.
5. **Dimensionality Reduction:** Eliminating irrelevant or redundant features.
6. **Data Splitting:** Dividing data into training, validation, and testing sets for ML model evaluation.

Effective data preprocessing ensures the accuracy and consistency of the dataset, leading to more reliable business intelligence and decision-making.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an open-ended, iterative approach to analyzing datasets to summarize their main characteristics, often using statistical graphics and other data visualization methods. It helps to identify patterns, spot anomalies, test hypotheses, and check assumptions.

Types of EDA:

1. **Univariate Non-Graphical:** Analyzing a single variable using statistical measures like mean, median, mode, variance, and standard deviation.
2. **Multivariate Non-Graphical:** Examining relationships between multiple variables using techniques like cross-tabulation, covariance, and correlation.
3. **Univariate Graphical:** Visualizing a single variable using plots such as histograms, box plots, and stem-and-leaf plots.
4. **Multivariate Graphical:** Visualizing relationships between multiple variables using scatter plots, heatmaps, and pair plots.

EDA is a crucial step in data analysis as it allows analysts to understand the data's structure and underlying patterns before applying formal modeling techniques.

1. (a) **Differentiate between covariance and correlation with example.** (8 marks)

Ans **(a) Covariance vs. Correlation**

Covariance and correlation are both statistical measures that describe the relationship between two variables, but they differ in scale and interpretability.

Aspect	Covariance	Correlation
Definition	Measures the directional relationship between two variables.	Measures both the strength and direction of the linear relationship between two variables.
Formula	cov(X,Y)=1n−1∑i=1n(Xi−X¯)(Yi−Y¯)\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})cov(X,Y)=n−11∑i=1n(Xi−X¯)(Yi−Y¯)	corr(X,Y)=cov(X,Y)σXσY\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}corr(X,Y)=σXσYcov(X,Y)
Range of Values	−∞-\infty−∞ to +∞+\infty+∞	−1-1−1 to +1+1+1
Units	Product of the units of X and Y	Unitless (dimensionless)
Interpretability	Harder to interpret due to dependence on units and scale	Easier to interpret; standardized measure

(b) **Explain in detail canonical correlation.** (7 marks)

Ans **(b) Canonical Correlation**

Canonical Correlation Analysis (CCA) is a multivariate statistical technique used to understand the relationship between two sets of variables by identifying linear combinations (canonical variates) of each set that are maximally correlated.Simon Fraser University

Key Concepts:

- **Canonical Variates**: Linear combinations of variables from each set that have the highest possible correlation.

- **Canonical Correlation Coefficients**: The correlation between each pair of canonical variates.

Steps in CCA:

1. **Standardization**: Center and scale the variables in both sets to have zero mean and unit variance.
2. **Compute Covariance Matrix**: Calculate the covariance matrix between the two sets of variables.
3. **Solve Eigenvalue Problem**: Solve the generalized eigenvalue problem to obtain canonical coefficients.
4. **Form Canonical Variates**: Construct the canonical variates using the canonical coefficients.
5. **Interpret Results**: Analyze the canonical correlation coefficients to understand the strength of the relationships between the sets.

Example:

Suppose we have two sets of variables:

- **Set X**: $[X_1, X_2, X_3]$ (e.g., physical measures)
- **Set Y**: $[Y_1, Y_2]$ (e.g., performance measures)

CCA will find linear combinations of X and Y, say $U = a_1X_1 + a_2X_2 + a_3X_3$ and $V = b_1Y_1 + b_2Y_2$, such that the correlation between U and V is maximized.

This technique is widely used in fields like psychology, ecology, and economics to explore complex relationships between variable sets.

Section B

4. **What is Hypothesis Testing? How is Hypothesis Testing Used in Data Science? Where and When to Use Hypothesis Test?** (15 marks)

Ans Hypothesis Testing in Data Science

Hypothesis testing is a fundamental statistical method used to assess the validity of a claim or assumption about a population based on sample data. It involves formulating two competing hypotheses:

- **Null Hypothesis (H_o): Assumes no effect or no difference exists.**
- **Alternative Hypothesis (H_1): Contradicts the null hypothesis, suggesting an effect or difference. The goal is to determine whether there is sufficient statistical evidence to reject the null hypothesis in favor of the alternative hypothesis.**

Application of Hypothesis Testing in Data Science

In data science, hypothesis testing is employed to make data-driven decisions and validate assumptions. Common applications include:

- **A/B Testing: Comparing two versions of a product or service to determine which performs better.**
- **Model Evaluation: Assessing whether a new model significantly improves over a baseline model.**
- **Feature Selection: Identifying which variables contribute significantly to the predictive power of a model.**
- **Anomaly Detection: Determining if a data point or pattern deviates significantly from the rest of the data.**

- **Quality Control: Monitoring processes to ensure consistent standards are met.**

These applications help data scientists draw conclusions about populations and make informed decisions based on sample data.

When to Use Hypothesis Testing

Hypothesis testing is appropriate when:

- **Making Comparisons: Evaluating differences between groups or conditions.**
- **Assessing Relationships: Determining associations between variables.**
- **Validating Assumptions: Testing preconceived notions or claims.**
- **Ensuring Quality: Monitoring processes for consistency and standards.**

It is particularly useful when decisions need to be based on data rather than subjective judgment.

5. Answer the following: (15)

(a) Is PCA or SVD better for dimensionality reduction?

Ans Dimensionality reduction is a crucial step in data preprocessing, especially when dealing with high-dimensional datasets. Two prominent techniques for this purpose are **Principal Component Analysis (PCA)** and **Singular Value Decomposition (SVD)**. While both aim to reduce the number of variables in a

dataset, they differ in methodology, application, and interpretability.

---

2. Mathematical Foundations

- **PCA**: PCA is a statistical method that transforms data into a set of orthogonal components that capture the maximum variance. It involves the eigenvalue decomposition of the covariance matrix of the data. The principal components are the eigenvectors corresponding to the largest eigenvalues, indicating the directions of maximum variance in the data.
- **SVD**: SVD is a matrix factorization technique that decomposes any m×nm \times nm×n matrix XXX into three matrices: X=UΣVTX = U \Sigma V^TX=UΣVT, where UUU and VVV are orthogonal matrices, and Σ\SigmaΣ is a diagonal matrix containing the singular values. The columns of UUU and VVV are the left and right singular vectors, respectively.

---

3. Relationship Between PCA and SVD

PCA can be computed using SVD. Given a centered data matrix XXX, performing SVD on XXX yields:

X=UΣVTX = U \Sigma V^TX=UΣVT

The principal components in PCA correspond to the columns of VVV, and the singular values in Σ\SigmaΣ relate to the variance captured by each principal component. Therefore, SVD provides a computational method to perform PCA.

---

4. Computational Considerations

- **PCA**: Computing PCA involves calculating the eigenvalues and eigenvectors of the covariance matrix, which can be computationally expensive, especially for large datasets.
- **SVD**: SVD is more versatile and can be applied directly to the data matrix without needing to compute the covariance matrix. It is particularly useful for sparse datasets and can be more efficient in certain scenarios.

---

5. Interpretability and Use Cases

- **PCA**: PCA provides components that are linear combinations of the original variables and are ordered by the amount of variance they capture. This makes PCA highly interpretable, especially in exploratory data analysis and visualization.
- **SVD**: SVD decomposes the data into orthogonal components but does not inherently provide components that maximize variance. While SVD is powerful for tasks like image compression and collaborative filtering, its components are less interpretable in terms of variance.

---

6. Summary Table

Feature	**PCA**	**SVD**
Purpose	Dimensionality reduction	Matrix factorization
Input	Covariance matrix	Data matrix
Output	Principal components	Singular vectors and values

Feature	PCA	SVD
Interpretability	High (variance explained)	Moderate (focus on decomposition)
Computational Cost	High (eigenvalue decomposition)	Variable (depends on implementation)
Best Suited For	Feature extraction, visualization	Sparse data, collaborative filtering

b) What is the difference between t-SNE and PCA for dimensionality reduction?

Ans **PCA vs. t-SNE: A Comparative Analysis**

1. Type of Technique

- **PCA**: A **linear** technique that transforms data into a new coordinate system by identifying directions (principal components) that maximize variance.
- **t-SNE**: A **non-linear** technique that focuses on preserving local structures by modeling pairwise similarities between data points and mapping them into a lower-dimensional space.

2. Structure Preservation

- **PCA**: Preserves **global structure**, meaning it maintains the overall variance and relationships between data points across the entire dataset.
- **t-SNE**: Preserves **local structure**, effectively capturing clusters and neighborhoods within the data, but may distort global relationships.

3. Interpretability

- **PCA**: Provides components that are linear combinations of the original variables, offering high interpretability.
- **t-SNE**: Results in a non-linear mapping, making the components less interpretable and primarily useful for visualization.

4. Computational Efficiency

- **PCA**: Computationally efficient, especially for large datasets, as it involves eigenvalue decomposition or singular value decomposition.
- **t-SNE**: Computationally intensive and less scalable for large datasets due to its iterative optimization process.

5. Determinism

- **PCA**: Deterministic; applying PCA to the same data will yield the same results each time.
- **t-SNE**: Stochastic; results can vary between runs due to its random initialization and optimization process.

6. Suitability for Data Types

- **PCA**: Best suited for data with linear relationships and when the goal is to reduce dimensionality while preserving variance.
- **t-SNE**: Ideal for visualizing high-dimensional data in 2D or 3D, particularly to identify clusters or patterns.

Summary Table

Feature	PCA	t-SNE
Type	Linear	Non-linear
Structure Preserved	Global	Local
Interpretability	High	Low
Computational Cost	Low	High
Deterministic	Yes	No
Best Use Case	Feature extraction, noise reduction	Data visualization, cluster identification

Section-C

5. What is the AR model in time series? How do we interpret Autoregressive Model? Explain its Process. (15)

Ans **Autoregressive (AR) Model in Time Series**

1. Definition

The **Autoregressive (AR) model** is a statistical model used to describe certain time-varying processes in which current values are expressed as a linear combination of their previous values and a stochastic error term. It assumes that past values have a direct influence on current values.

Mathematically, an AR model of order ppp, denoted as AR(ppp), is represented as:

$Xt=\phi1Xt−1+\phi2Xt−2+⋯+\phipXt−p+\varepsilon tX_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \dots + \phi_p X_{t-p} + \varepsilon_tXt=\phi1Xt−1+\phi2Xt−2+⋯+\phipXt−p+\varepsilon t$

Where:

- XtX_tXt: Value at time ttt
- $\phi1,\phi2,…,\phip\phi_1, \phi_2, \dots, \phi_p\phi1,\phi2,…,\phip$: Parameters of the model
- $\varepsilon t\varepsilon_t\varepsilon t$: White noise error term at time ttt

2. Interpretation

- **Lagged Dependence**: The model captures the dependency of the current value on its past values. For instance, in an AR(1) model, the current value depends on its immediate past value.
- **Stationarity**: For the AR model to be valid, the time series must be stationary, meaning its statistical properties do not change over time. This implies constant mean, variance, and autocorrelation over time.
- **Parameter Significance**: The parameters $\phi1,\phi2,…,\phip\phi_1, \phi_2, \dots, \phi_p\phi1,\phi2,…,\phip$ represent the strength and direction of the influence of past values on the current value. A positive $\phi\phi\phi$ indicates a direct relationship, while a negative $\phi\phi\phi$ indicates an inverse relationship.

3. Process of AR Model

a. Stationarity Check

Before fitting an AR model, it's crucial to ensure that the time series is stationary. Techniques like the Augmented Dickey-Fuller (ADF) test can be employed to test for stationarity. If the series is non-stationary, differencing or transformation methods may be applied.

b. Model Identification

The order ppp of the AR model is determined using:

- **Autocorrelation Function (ACF)**: Measures the correlation between the series and its lagged versions.
- **Partial Autocorrelation Function (PACF)**: Measures the correlation between the series and its lagged versions after removing the effects of shorter lags.

A significant spike in the PACF at lag ppp suggests an AR(ppp) model.

c. Parameter Estimation

Once the model order is identified, the parameters ϕ1,ϕ2,…,ϕp\phi_1, \phi_2, \dots, \phi_pϕ1,ϕ2,…,ϕp are estimated using methods like:

- **Ordinary Least Squares (OLS)**: Minimizes the sum of squared differences between observed and predicted values.
- **Maximum Likelihood Estimation (MLE)**: Maximizes the likelihood of observing the given data under the model.

d. Model Diagnostics

After fitting the model, it's essential to check the residuals (errors) to ensure they resemble white noise:

- **Autocorrelation of Residuals**: Use the ACF of residuals to check for any remaining patterns.
- **Ljung-Box Test**: A statistical test to check if there are significant autocorrelations at any lag.
- **Durbin-Watson Statistic**: Tests for the presence of autocorrelation in residuals. A value around 2 indicates no autocorrelation .

e. Forecasting

Once the model is validated, it can be used to forecast future values. The forecasted value at time t+1t+1t+1 is:

X^t+1=ϕ1Xt+ϕ2Xt−1+⋯+ϕpXt−p+1\hat{X}_{t+1} = \phi_1 X_t + \phi_2 X_{t-1} + \dots + \phi_p X_{t-p+1}X^t+1=ϕ1Xt+ϕ2Xt−1+⋯+ϕp Xt−p+1

4. Practical Considerations

- **Model Order Selection**: The choice of ppp is critical. Overfitting can occur with a high ppp, while underfitting can occur with a low ppp. Information criteria like AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) can aid in selecting the optimal order.
- **Model Limitations**: The AR model assumes linear relationships and may not capture complex nonlinear patterns in the data. In such cases, other models like ARMA (Autoregressive Moving Average) or ARIMA (Autoregressive Integrated Moving Average) might be more appropriate.

---

6. Describe the Autoregressive Integrated Moving Average (ARIMA) model. What are the three main components of the ARIMA model, and how are they combined to model a time series? (15)

ANS 📘 **What is the ARIMA Model?**

The **ARIMA** model is a statistical approach used for analyzing and forecasting time series data. It is particularly effective for datasets

that exhibit patterns such as trends or cycles but lack seasonality. ARIMA combines three key components:

- **Autoregressive (AR)**: This component models the current value of the series as a linear combination of its previous values. It captures the relationship between an observation and a number of lagged observations.
- **Integrated (I)**: Differencing the raw observations to make the time series stationary, i.e., to remove trends or cycles. This step is crucial because many statistical forecasting methods assume stationarity.
- **Moving Average (MA)**: This component models the relationship between an observation and a residual error from a moving average model applied to lagged observations.

The ARIMA model is typically denoted as **ARIMA(p, d, q)**, where:

- **p** = the number of lag observations included in the model (AR term)
- **d** = the number of times that the raw observations are differenced (I term)
- **q** = the size of the moving average window (MA term)

🧩 Components of ARIMA

1. Autoregressive (AR) Component

The AR part of the model indicates that the evolving variable of interest is regressed on its prior values. The number of lag observations included in the model is denoted as **p**.

2. Integrated (I) Component

The I part involves differencing the raw observations to make the time series stationary. The number of times that the raw observations are differenced is denoted as **d**.

3. Moving Average (MA) Component

The MA part models the relationship between an observation and a residual error from a moving average model applied to lagged observations. The size of the moving average window is denoted as **q**.

🔁 How ARIMA Models Time Series

1. **Stationarity Check**: First, assess if the time series is stationary. If not, apply differencing (the I component) to achieve stationarity.
2. **Model Identification**: Determine the values of p, d, and q using plots like the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF), or by using criteria such as the Akaike Information Criterion (AIC).
3. **Parameter Estimation**: Estimate the parameters of the AR and MA components using methods like Maximum Likelihood Estimation (MLE).
4. **Model Diagnostics**: Check the residuals of the model to ensure they resemble white noise, indicating that the model has adequately captured the underlying patterns in the data.
5. **Forecasting**: Use the fitted ARIMA model to make forecasts about future values of the time series.

Section-D

7. Compare the efficiency and accuracy of bootstrapping and Monte Carlo methods in estimating statistical parameters or solving numerical problems. Under what circumstances would you choose one method over the other, and why? (15)

Ans Bootstrapping and Monte Carlo simulations are both resampling techniques used to estimate statistical parameters and solve numerical problems. While they share similarities, they differ in their methodologies, assumptions, and applications.

🧩 Bootstrapping

Definition: Bootstrapping is a non-parametric statistical method that involves repeatedly sampling with replacement from an observed dataset to estimate the sampling distribution of a statistic.

Key Characteristics:

- **Assumptions**: Minimal; relies on the observed data without assuming a specific underlying distribution.
- **Data Dependency**: Utilizes the original dataset, making it suitable for small or complex datasets where traditional parametric methods may not be applicable.Wikipedia
- **Computational Efficiency**: Generally less computationally intensive compared to Monte Carlo simulations.

Limitations:

- **Dependence on Original Data**: The accuracy of bootstrap estimates depends on the representativeness of the original dataset.
- **Not Suitable for All Data Types**: May not perform well with data exhibiting strong temporal dependencies or spatial correlations.

🎲 Monte Carlo Simulation

Definition: Monte Carlo simulations involve using random sampling to obtain numerical results, often to estimate complex integrals or solve problems with significant uncertainty.

Key Characteristics:

- **Assumptions**: Requires assumptions about the underlying probability distributions of input variables.
- **Flexibility**: Highly flexible; can model complex systems and processes across various domains.
- **Computational Intensity**: Can be computationally intensive, especially for high-dimensional problems.

Limitations:

- **Model Dependency**: The accuracy of results is highly dependent on the correctness of the model and assumptions used.
- **Computational Cost**: High computational cost for problems requiring a large number of simulations.

⚖️ Comparison Table

Feature	Bootstrapping	Monte Carlo Simulation
Assumptions	Minimal (non-parametric)	Requires distributional assumptions
Data Requirements	Relies on observed data	Requires model inputs and assumptions
Computational Cost	Generally lower	Can be high, especially for complex models
Flexibility	Limited to resampling from existing data	Highly flexible; can model complex systems
Accuracy Dependence	Depends on representativeness of data	Depends on model correctness and assumptions
Best Use Cases	Estimating confidence intervals, bias, variance	Estimating integrals, simulating complex systems

8 Explain the following: (15)

(a) EHR data

(b) Price optimisation in retail

(c) Demand forecasting

Ans **(a) Electronic Health Record (EHR) Data**

Definition: An Electronic Health Record (EHR) is a digital version of a patient's paper chart, encompassing a comprehensive collection of health information. EHRs are designed to be shared across different healthcare settings, facilitating coordinated and efficient care.

Key Components:

- **Patient Demographics**: Includes personal information such as age, gender, and contact details.
- **Medical History**: Documents past illnesses, surgeries, and family medical history.
- **Medications and Allergies**: Lists current prescriptions and known allergies.
- **Immunization Status**: Records of vaccinations received.
- **Laboratory and Test Results**: Includes blood tests, imaging results, and other diagnostic information.
- **Vital Signs**: Data such as blood pressure, heart rate, and temperature.
- **Progress Notes**: Clinician's observations and treatment plans.

Benefits:

- **Improved Patient Care**: EHRs provide healthcare providers with accurate, up-to-date patient information, leading to better decision-making and reduced medical errors.
- **Enhanced Coordination**: Facilitates sharing of patient data among specialists, ensuring comprehensive care.
- **Efficiency Gains**: Reduces paperwork, streamlines workflows, and minimizes duplication of tests.

Challenges:

- **Interoperability Issues**: Different EHR systems may not communicate effectively, hindering data exchange.
- **Data Security Concerns**: Sensitive health information is vulnerable to cyber threats if not properly protected. **Costs**: Transitioning to EHR systems can be expensive and time-consuming for healthcare providers.

(b) Price Optimization in Retail

Definition: Price optimization in retail involves using data analytics to determine the most effective pricing strategies that maximize revenue and profitability while considering factors like demand elasticity, competition, and customer behavior.

Key Strategies:

- **Dynamic Pricing**: Adjusting prices in real-time based on market demand, competitor pricing, and other external factors.
- **Segmentation-Based Pricing**: Setting different prices for different customer segments based on purchasing behavior and willingness to pay.
- **Promotional Pricing**: Offering discounts or special deals to stimulate sales during specific periods.

Benefits:

- **Revenue Maximization**: Identifies optimal price points that balance demand and profitability.
- **Competitive Advantage**: Allows retailers to respond swiftly to market changes and competitor actions.

- **Customer Satisfaction**: Ensures customers perceive value in the pricing, enhancing loyalty.

Challenges:

- **Data Dependency**: Requires accurate and comprehensive data to make informed pricing decisions.
- **Customer Perception**: Frequent price changes can lead to customer dissatisfaction if not managed transparently.
- **Implementation Complexity**: Integrating price optimization tools into existing retail systems can be complex.

(c) Demand Forecasting

Definition: Demand forecasting is the process of predicting future customer demand for products or services using historical data, market analysis, and statistical tools.

Methods:

- **Qualitative Methods**: Based on expert judgment and market research, used when historical data is limited.
- **Quantitative Methods**: Utilize historical sales data and statistical models to predict future demand.

Benefits:

- **Inventory Optimization**: Helps in maintaining optimal stock levels, reducing overstocking and stockouts.
- **Cost Efficiency**: Improves resource allocation and reduces operational costs.
- **Strategic Planning**: Informs marketing, production, and procurement strategies.

Challenges:

- **Data Quality**: Accurate forecasts depend on the quality and completeness of historical data.
- **Market Volatility**: Sudden market changes can render forecasts inaccurate.
- **Model Complexity**: Developing and maintaining forecasting models can be resource-intensive.