Pimpri Chinchwad College of Engineering, Nigdi, Pune
Department of Information Technology

TYBTECH BIT5508 Foundations of Data Science

Question Bank
AY 2024-25

UNIT –I
Introduction to Data Science

Sr. No	Questions	Bloom's Level	CO Mapped
1	Model the primary differences between structured, semi-structured, and unstructured data? Provide examples of each and discuss how they are typically stored and processed.	L3	CO1
2	In the context of big data analytics, explain the concept of "data variety." How does the presence of different data types (e.g., text, images, audio) impact the selection of analytical techniques and tools?	L2	CO1
3	Illustrate the role of metadata in big data systems. How does metadata help in managing and interpreting different types of data within a large-scale data environment?	L3	CO1
4	Implement data science techniques be used to convert unstructured data into a structured format suitable for analysis? Discuss at least two methods and their implications for data processing.	L3	CO1
5	Interpret challenges arise when integrating data from multiple sources with different types of data (e.g., sensor data, social media data, transactional data)? How can data scientists address these challenges to ensure cohesive data analysis?	L2	CO1
6	Explain the fundamental definition of Machine Learning and how it differentiates from traditional programming methods.	L2	CO1
7	Illustrate the potential risks and challenges associated with data collection in a large-scale survey, and how can these be mitigated?	L3	CO1
8	Describe a scenario where data collection biases could significantly impact the results of a data science project. How would you address these biases?	L3	CO1
9	How can you integrate data collected from multiple sources with varying formats and structures into a cohesive dataset?	L3	CO1
10	Discuss the ethical considerations and privacy concerns when collecting sensitive personal data for a data science project.	L3	CO1
11	Demonstrate the difference between training error, validation error, and test error. Why is it important to evaluate a model on a separate test set, and how does it contribute to assessing model performance?	L3	CO1
12	Illustrate some advanced techniques for evaluating model performance on imbalanced datasets, such as precision-recall curves or stratified sampling? How do these techniques provide a more comprehensive assessment compared to traditional metrics?	L3	CO1
13	Discuss the trade-offs between using a heat map versus a scatter plot for visualizing large datasets. In what scenarios might one be preferred over the other?	L2	CO1
14	Explain how interactive dashboards can enhance data visualization. What are the key elements to consider when designing an interactive dashboard?	L3	CO1
15	Describe the principles of effective color use in data visualization. How can poor color choices impact data interpretation and decision-making?	L2	CO1

16	Demonstrate how do you determine the most appropriate type of chart or graph for visualizing time-series data? Provide examples of different types and their applications.	**L3**	**CO1**
17	Discuss the role of data aggregation in data visualization. How does aggregating data affect the clarity and usability of the visualization?	**L2**	**CO1**
18	Evaluate the effectiveness of using 3D charts for complex data sets. What are the potential advantages and disadvantages compared to 2D charts?	**L3**	**CO1**
19	Investigate How can data normalization techniques impact the visualization of comparative data? Provide examples of techniques used to normalize data and their effect on visualization.	**L4**	**CO1**
20	Analyze the use of heat maps in identifying patterns and anomalies in large datasets. What are some best practices for creating and interpreting heat maps?	**L4**	**CO1**
21	Explain the concept of visual encoding in data visualization. How does visual encoding affect the accuracy and ease of interpreting data?	**L4**	**CO1**
22	Discuss the impact of data visualization on storytelling and communication. How can visualizations be used effectively to convey a narrative or make a persuasive argument?	**L4**	**CO1**
23	Using a dataset of sales figures across different regions, create a bar chart to visualize the sales distribution by region. Explain how this visualization helps in identifying patterns or trends in sales performance and suggest potential business decisions that could be informed by this data.	**L3**	**CO1**
24	Given a dataset containing information on customer demographics and purchase behavior, use a scatter plot to visualize the relationship between age and spending amount. Describe how this visualization can help in understanding customer behavior and identifying target segments for marketing.	**L3**	**CO1**
25	Create a line chart to visualize the trend of website traffic over a year. Analyze the chart to identify peak periods of traffic and discuss potential factors that could have influenced these trends. How can this information guide future marketing strategies	**L3**	**CO1**
26	Using a dataset with multiple categorical variables, apply a stacked bar chart to visualize the distribution of categories within each variable. Explain how this visualization provides insights into the relationships between the categories and suggest how these insights could inform business decisions.	**L3**	**CO1**
27	Given a dataset of test scores from students in different classes, create a box plot to visualize the distribution of scores. Explain how this visualization helps in identifying outliers and understanding the spread of scores within each class. Discuss how this information could be used to improve educational outcomes.	**L3**	**CO1**
28	Given a dataset containing customer purchase histories, outline the steps you would take to clean and preprocess the data before analysis. Explain how each step (e.g., handling missing values, data normalization, and encoding categorical variables) contributes to the overall quality of the analysis.	**L3**	**CO1**
29	Using a dataset related to housing prices, apply exploratory data analysis (EDA) techniques to summarize the main characteristics of the data. Create visualizations (e.g., histograms, scatter plots) to identify relationships between features and the target variable. Discuss the insights gained from your analysis.	**L3**	**CO1**
30	After developing a machine learning model to predict customer churn, describe the steps you would take to evaluate its performance. Include metrics you would use (e.g., accuracy, precision, recall) and explain why these metrics are important for assessing model effectiveness."	**L3**	**CO1**
31	Explain Data Science process with suitable example	**L3**	**CO1**
32	Explain why the data is divided in to training and testing sets	**L3**	**CO1**

UNIT –II

Mathematical foundation for Data Science

Sr. No	Questions	Bloom's Level	CO Mapped
	Interpret the Mean: Given a dataset representing the scores of students in a math exam, calculate the mean score. Explain what this value indicates about the overall performance of the students.	L3	CO2
	Analyze the Median: For a dataset of ages of participants in a community event, calculate the median age. Discuss how the median provides a better understanding of the age distribution compared to the mean in this context.	L3	CO2
	Evaluate the Mode: Examine the frequency distribution of colors preferred by a group of respondents. Identify the mode and explain its significance in understanding group preferences.	L3	CO2
	Explain the concept of correlation. What does a correlation coefficient indicate about the relationship between two variables?	L2	CO2
	If a financial analyst observes a high positive covariance between the returns of two stocks, how can this information be applied to a portfolio management strategy?	L3	CO2
	Given two sets of data representing the monthly sales of two products, calculate the covariance. What does this value suggest about the relationship between the sales of the two products?	L3	CO2
	Discuss the significance of the units of covariance and how they can affect the interpretation of its value.	L3	CO2
	What does a positive covariance indicate about two variables, and how does it relate to their movement in relation to each other	L2	CO2
	Define covariance. How does it differ from correlation in measuring the relationship between two variables?	L2	CO2
	What are the differences between positive, negative, and zero correlation? Provide examples of each.	L2	CO2
	Describe how a scatter plot can be used to visualize the correlation between two variables.	L2	CO2
	If a researcher finds a strong negative correlation between hours spent on social media and academic performance, how might they apply this finding to develop strategies for improving student performance?	L3	CO2
	Given a dataset of students' study hours and their exam scores, calculate the correlation coefficient and interpret its meaning in context.	L3	CO2
	Apply Range Calculation: A teacher recorded the number of books read by each student in a class. Calculate the range of this dataset and describe what this range reveals about the variability in reading habits among the students.	L3	CO2
	Assess Standard Deviation: Given the daily temperatures recorded over a week, compute the standard deviation. Discuss how this measure helps in understanding the consistency of temperatures during that week.	L3	CO2
	Construct a Box Plot: Create a box plot for a set of exam scores. Use the plot to explain the interquartile range and what it reveals about the score distribution.	L3	CO2
	Interpret Quartiles: For a dataset representing household incomes, calculate the first and third quartiles. Explain how these quartiles can help identify income disparities within the dataset.	L3	CO2
	Utilize Skewness: Given a dataset of monthly expenses for households, analyze the skewness of the data. Discuss what the direction of the skewness indicates	L3	CO2

	about spending behavior.		
1.	Apply Percentiles: Calculate the 90th percentile of a dataset representing the number of hours students spend studying each week. Discuss the implications of this percentile for understanding study habits among high-performing students.	**L3**	**CO2**
2.	Evaluate Data Visualization: Given a histogram of sales data for a product over several months, interpret the shape of the histogram. Discuss what insights can be drawn regarding trends in product sales from the visualization.	**L3**	**CO2**
3.	Interpret the Mean: Given a dataset representing the scores of students in a math exam, calculate the mean score. Explain what this value indicates about the overall performance of the students.	**L3**	**CO2**
4.	Analyze the Median: For a dataset of ages of participants in a community event, calculate the median age. Discuss how the median provides a better understanding of the age distribution compared to the mean in this context.	**L3**	**CO2**
5.	Evaluate the Mode: Examine the frequency distribution of colors preferred by a group of respondents. Identify the mode and explain its significance in understanding group preferences.	**L3**	**CO2**
6.	Apply Range Calculation: A teacher recorded the number of books read by each student in a class. Calculate the range of this dataset and describe what this range reveals about the variability in reading habits among the students.	**L3**	**CO2**
7.	Assess Standard Deviation: Given the daily temperatures recorded over a week, compute the standard deviation. Discuss how this measure helps in understanding the consistency of temperatures during that week.	**L3**	**CO2**
8.	Construct a Box Plot: Create a box plot for a set of exam scores. Use the plot to explain the interquartile range and what it reveals about the score distribution.	**L3**	**CO2**
9.	Interpret Quartiles: For a dataset representing household incomes, calculate the first and third quartiles. Explain how these quartiles can help identify income disparities within the dataset.	**L3**	**CO2**
10.	Utilize Skewness: Given a dataset of monthly expenses for households, analyze the skewness of the data. Discuss what the direction of the skewness indicates about spending behavior.	**L3**	**CO2**
11.	Apply Percentiles: Calculate the 90th percentile of a dataset representing the number of hours students spend studying each week. Discuss the implications of this percentile for understanding study habits among high-performing students.	**L3**	**CO2**
12.	Evaluate Data Visualization: Given a histogram of sales data for a product over several months, interpret the shape of the histogram. Discuss what insights can be drawn regarding trends in product sales from the visualization.	**L3**	**CO2**
13.	Given two events, A and B, where $P(A) = 0.4$, $P(B) = 0.5$, and $P(A \cap B) = 0.2$, determine whether A and B are independent. Provide a detailed explanation of your calculations and reasoning.	**L3**	**CO2**
14.	Evaluate Independence in Context: Consider a study that examines the relationship between smoking and lung cancer. Analyze the data provided to determine if smoking and lung cancer are independent events. Discuss the implications of your findings on public health	**L3**	**CO2**

	initiatives.		
15.	Investigate Real-World Dependencies: Examine a dataset showing the correlation between students' study hours and their final exam scores. Analyze whether these two variables can be considered independent or dependent. Justify your analysis with statistical evidence.	**L3**	**CO2**
16.	Compare Different Scenarios: Analyze two experiments: one where the outcome of rolling a die is independent of flipping a coin, and another where the outcome of drawing a card from a deck is dependent on the first draw. Compare and contrast the nature of dependence in these scenarios, and discuss how this affects the outcomes.	**L3**	**CO2**
17.	Evaluate the Impact of Independence Assumptions: In a marketing analysis, a company assumes that customer purchase decisions are independent of their age group. Analyze the potential risks of this assumption. Discuss how failure to recognize dependence might lead to ineffective marketing strategies.	**L4**	**CO2**
18.	Given two events, A and B, where P(A) = 0.4, P(B) = 0.5, and P(A ∩ B) = 0.2, determine whether A and B are independent. Provide a detailed explanation of your calculations and reasoning.	**L4**	**CO2**
19.	AnalyzeConditional Relationships: Given a medical test with a 95% sensitivity and a 90% specificity for detecting a disease that affects 1% of the population, analyze the likelihood that a person who tests positive actually has the disease. Show your calculations clearly.	**L4**	**CO2**
20.	Evaluate Bayes's Theorem Application: In a survey, 70% of respondents prefer Brand A over Brand B. If 80% of Brand A consumers are satisfied with their choice while only 40% of Brand B consumers express satisfaction, analyze the overall satisfaction rate of a randomly selected consumer. Use Bayes's Theorem to justify your findings.	**L4**	**CO2**
21.	Investigate Misinterpretations: Analyze a scenario where a lawyer uses conditional probability to argue that a defendant is innocent based on a low probability of a similar crime occurring. Discuss the potential pitfalls in this reasoning, considering prior probabilities and how they affect conclusions.	**L4**	**CO2**
22.	Compare Two Conditional Probabilities: You have data on two different events: the likelihood of passing an exam given that a student attended extra classes versus the likelihood of passing without attending. Analyze these probabilities and discuss what factors could influence the differences observed.	**L4**	**CO2**
23.	Examine Sequential Events: A bag contains 3 red balls and 2 blue balls. If two balls are drawn sequentially without replacement, analyze the probability that the first ball is red and the second is blue. Clearly outline your reasoning and calculations.	**L4**	**CO2**
24.	Apply Bayes's Theorem in Decision Making: A factory produces a product with a 95% defect-free rate. If a random sample of 10 products shows 2 defects, analyze the probability that the production line is functioning correctly, using Bayes's Theorem. Discuss how this information could impact quality control decisions.	**L4**	**CO2**
25.	Analyze Conditional Relationships: Given a medical test with a 95% sensitivity and a 90% specificity for detecting a disease that affects 1% of the population, analyze the likelihood that a person who tests positive	**L4**	**CO2**

	actually has the disease. Show your calculations clearly.		
26.	Evaluate Assumptions in Conditional Probability: Assess the assumptions made when applying conditional probability in a given scenario, such as predicting the likelihood of weather events based on historical data. Discuss the potential consequences of these assumptions on decision-making.	**L5**	**CO2**
27.	Critique Bayes's Theorem Application: Critically evaluate the use of Bayes's Theorem in a clinical trial where the prior probabilities of treatment effectiveness are not well established. Discuss how this uncertainty can impact the reliability of the conclusions drawn.	**L5**	**CO2**
28.	Assess the Role of Conditional Probability in Risk Management: In a business context, evaluate how conditional probabilities can be used to assess risks in investment decisions. Discuss the potential benefits and drawbacks of relying on these probabilities.	**L5**	**CO2**
29.	Evaluate the Impact of Sample Size: In the context of a study evaluating the effectiveness of a new drug, assess how the size of the sample affects the conditional probabilities calculated. Discuss how a small sample size might lead to misleading conclusions.	**L5**	**CO2**
30.	Review Ethical Implications of Misusing Bayes's Theorem: Analyze a case where Bayes's Theorem is misapplied in legal contexts, leading to wrongful convictions or acquittals. Discuss the ethical implications and the importance of correct probability assessment in judicial processes.	**L5**	**CO2**
31.	Evaluate Assumptions in Conditional Probability: Assess the assumptions made when applying conditional probability in a given scenario, such as predicting the likelihood of weather events based on historical data. Discuss the potential consequences of these assumptions on decision-making.	**L5**	**CO2**
32.	Explain the Central Limit Theorem: Describe the Central Limit Theorem and its significance in statistics. Provide an example of its application in a real-world scenario.	**L3**	**CO2**
33.	Apply CLT to Sample Means: Given a population with a known mean and standard deviation, calculate the expected mean and standard deviation of the sampling distribution of the sample mean for a specified sample size.	**L3**	**CO2**
34.	Use the CLT for Approximation: Apply the Central Limit Theorem to approximate the distribution of sample means when sampling from a population that is not normally distributed.	**L3**	**CO2**
35.	Interpret Sampling Distributions: Given a sample mean from a population, explain how the Central Limit Theorem allows for inferences about the population mean.	**L3**	**CO2**
36.	Demonstrate CLT with a Simulation: Conduct a simulation to demonstrate the Central Limit Theorem. Describe the process and analyze the results in terms of convergence to a normal distribution.	**L3**	**CO2**
37.	Explain the Central Limit Theorem: Describe the Central Limit Theorem and its significance in statistics. Provide an example of its application in a real-world scenario.	**L3**	**CO2**
38.	Analyze the Implications of the CLT: Examine how the Central Limit Theorem justifies the use of normal approximation in hypothesis testing and confidence intervals, even for non-normally distributed populations.	**L4**	**CO2**

39.	Investigate Sample Size Effects: Analyze how varying sample sizes affect the applicability of the Central Limit Theorem. Discuss the minimum sample size required for normal approximation in different contexts.	**L4**	**CO2**
40.	Evaluate Limitations of the CLT: Critically analyze the limitations of the Central Limit Theorem in practical applications, particularly regarding skewed distributions and outliers.	**L4**	**CO2**
41.	Explore the Role of Independence: Investigate how the assumption of independence among samples affects the validity of the Central Limit Theorem. Provide examples to support your analysis.	**L4**	**CO2**
42.	Compare the CLT with Other Theorems: Compare the Central Limit Theorem with other statistical theorems (e.g., law of large numbers) and analyze their implications in statistical inference.	**L4**	**CO2**
43.	Analyze the Implications of the CLT: Examine how the Central Limit Theorem justifies the use of normal approximation in hypothesis testing and confidence intervals, even for non-normally distributed populations.	**L4**	**CO2**
44.	Evaluate the Practicality of the CLT: Critically assess the practical implications of the Central Limit Theorem in statistical research. Discuss potential biases that may arise from misapplying the theorem.	**L5**	**CO2**
45.	Assess the Role of the CLT in Data Science: Evaluate the significance of the Central Limit Theorem in modern data science practices, especially in machine learning and predictive modeling.	**L5**	**CO2**
46.	Review Ethical Considerations: Analyze the ethical implications of using the Central Limit Theorem in data analysis, particularly when it comes to generalizing results to broader populations.	**L5**	**CO2**
47.	Examine Case Studies: Critique case studies that rely on the Central Limit Theorem for inference. Discuss the strengths and weaknesses of their methodologies and conclusions.	**L5**	**CO2**
48.	Critique Research that Misapplies the CLT: Analyze a research study that incorrectly applies the Central Limit Theorem. Discuss how this affects the validity of its findings and potential consequences.	**L5**	**CO2**
49.	Identify Properties of the Normal Distribution: Given a dataset, identify whether it approximates a normal distribution based on its shape and calculate its mean and standard deviation.	**L3**	**CO2**
50.	Calculate Probabilities Using Z-Scores: Given a normally distributed variable with a specified mean and standard deviation, calculate the probability that a randomly selected value falls above a certain threshold using Z-scores.	**L3**	**CO2**
51.	Apply the Empirical Rule: Use the empirical rule (68-95-99.7 rule) to explain the distribution of data points in a normally distributed dataset.	**L3**	**CO2**
52.	Interpret Confidence Intervals: Calculate a confidence interval for the mean of a normally distributed dataset. Interpret what this interval means in a practical context.	**L3**	**CO2**
53.	Use Normal Distribution in Quality Control: Apply the concept of normal distribution to determine acceptable limits for a manufacturing process based on given specifications.	**L3**	**CO2**

54.	Identify Properties of the Normal Distribution: Given a dataset, identify whether it approximates a normal distribution based on its shape and calculate its mean and standard deviation.	**L3**	**CO2**
55.	Analyze Deviations from Normality: Examine a dataset and determine if it significantly deviates from a normal distribution. Discuss the implications of these deviations for statistical analysis.	**L4**	**CO2**
56.	Investigate Real-World Applications: Analyze how the normal distribution is used in various fields, such as psychology or economics, and discuss its relevance in those contexts.	**L4**	**CO2**
57.	Evaluate Transformation Effects: Investigate the effects of data transformation (e.g., log transformation) on a non-normally distributed dataset to assess normality. Discuss the outcomes.	**L4**	**CO2**
58.	Compare Normal and Other Distributions: Compare the characteristics and applications of the normal distribution with another distribution (e.g., binomial or Poisson). Discuss their appropriateness in various contexts.	**L4**	**CO2**
59.	Examine Statistical Methods for Normality: Analyze the methods used to test for normality in a dataset. Discuss their effectiveness and potential limitations.	**L4**	**CO2**
60.	Analyze Deviations from Normality: Examine a dataset and determine if it significantly deviates from a normal distribution. Discuss the implications of these deviations for statistical analysis.	**L4**	**CO2**
61.	Investigate Real-World Applications: Analyze how the normal distribution is used in various fields, such as psychology or economics, and discuss its relevance in those contexts.	**L4**	**CO2**
62.	Evaluate Assumptions of Normality: Critically evaluate the assumptions made when applying statistical tests that rely on normality, and discuss the implications of violating these assumptions.	**L5**	**CO2**
63.	Assess Impact of Non-Normality: Analyze the impact of non-normality on the results of statistical analyses in a given study. Discuss how this can lead to incorrect conclusions.	**L5**	**CO2**
64.	Review Ethical Considerations: Evaluate the ethical implications of using normal distribution assumptions in research, particularly in sensitive areas such as healthcare or education.	**L5**	**CO2**
65.	Examine Policy Decisions Based on Normal Distribution: Critically assess a policy decision that relies on normal distribution analysis. Discuss the robustness of the conclusions and any potential biases.	**L5**	**CO2**
66.	Critique Research Studies Using Normality Assumptions: Analyze a research study that employs normal distribution assumptions. Discuss strengths, weaknesses, and the validity of the conclusions drawn.	**L5**	**CO2**
67.	Identify Types of Random Variables: Given a scenario involving the number of heads obtained when flipping three coins, identify whether the random variable is discrete or continuous. Justify your answer.	**L3**	**CO2**
68.	Calculate Expected Value: A game involves rolling a fair six-sided die and winning the number of dollars equal to the number shown. Calculate the expected value of the winnings.	**L3**	**CO2**
69.	Apply Probability Mass Function: For a discrete random variable representing the number of defective items in a batch of 10, use the provided probability mass function to determine the probability of	**L3**	**CO2**

	finding exactly 3 defective items.		
70.	Use Cumulative Distribution Function: Given a cumulative distribution function (CDF) for a discrete random variable, calculate the probability that the variable takes a value less than or equal to a certain number.	**L3**	**CO2**
71.	Interpret Random Variable Outcomes: If a random variable represents the height of students in a classroom, describe how this variable can be used to make predictions about the overall height distribution in a larger population.	**L3**	**CO2**
72.	Identify Types of Random Variables: Given a scenario involving the number of heads obtained when flipping three coins, identify whether the random variable is discrete or continuous. Justify your answer.	**L3**	**CO2**
8	Given the following dataset, calculate the IQR and determine any outliers: [12, 15, 14, 22, 18, 30, 35, 27, 29, 31].Walk through the steps of finding Q1, Q3, and the IQR, and use the IQR method to detect outliers.	**L3**	**CO2**
	Explain the process of calculating the first quartile (Q1), third quartile (Q3), and how the IQR is derived.	**L3**	**CO2**
	How can the IQR method be used to detect outliers in a dataset?** - Describe the formula for identifying outliers using IQR and how to apply it to both the lower and upper bounds of a dataset.	**L3**	**CO2**
	What are the advantages of using the IQR method for outlier detection compared to other methods like z-scores? - Discuss the robustness of IQR against skewed data and the benefits of using it for non-normal distributions.	**L3**	**CO2**
	How would you handle outliers detected using the IQR method in real-world datasets? Explore different approaches such as removing outliers, transforming the data, or using alternative techniques, depending on the context of the analysis.	**L3**	**CO2**

UNIT –III

Data Pre-Processing

Sr. No	Questions	Bloom's Level	CO Mapped
1.	Identify and remove any outliers from a given dataset using Z-score or IQR (Interquartile Range) methods. Show your process and the resulting cleaned dataset.	**L3**	**CO3**
2.	Apply a data transformation technique (e.g., log transformation or Box-Cox transformation) to correct skewness in a given dataset. Demonstrate the steps and explain the effect of the transformation on the data distribution.	**L3**	**CO3**
3.	Apply normalization and standardization techniques to a dataset. Show the steps and the resulting transformation for at least two features, and explain when you would choose one technique over the other	**L3**	**CO3**
4.	Given a dataset with missing values in multiple columns, demonstrate how you would handle these missing values using appropriate techniques. Explain your reasoning behind choosing each technique	**L3**	**CO3**
5.	Using a sample dataset, demonstrate how you would handle categorical variables with more than two categories using encoding techniques such as one-hot encoding or label encoding. Explain your choice of technique.	**L3**	**CO3**
6.	Examine a given dataset and identify any patterns or trends using summary statistics (e.g., mean, median, standard deviation). Explain how these statistics help in understanding the data.	**L3**	**CO3**
7.	Segment the data based on a categorical variable (e.g., gender, region) and compare the mean and variance of a numerical variable across different groups. Explain how the segmentation helps in understanding the differences within the data.	**L3**	**CO3**
8.	Given a dataset with missing values in both numerical and categorical columns, apply appropriate imputation techniques (e.g., mean, median, mode, or most frequent) to handle the missing data. Explain the rationale for choosing each technique.	**L3**	**CO3**
9.	Using a dataset where certain rows have multiple missing values, apply the method of row deletion (listwise or pairwise deletion). After cleaning, explain the impact of this technique on the dataset size and analysis results.	**L3**	**CO3**
10.	Identify missing patterns in a dataset (e.g., Missing Completely at Random, Missing at Random, or Missing Not at Random). Apply a suitable strategy for each pattern and demonstrate how this improves the dataset's completeness and integrity.	**L3**	**CO3**
11.	Implement interpolation methods (e.g., linear, polynomial, or spline interpolation) to fill in missing values for a time-series dataset. Compare the results with other imputation techniques and explain which method yields better results for time-series data.	**L3**	**CO3**
12.	Using a dataset that contains redundant symbols (e.g., currency symbols, percentage signs), apply appropriate formatting to remove these symbols and convert the data into plain numerical values. Explain how this formatting makes the data more suitable for statistical analysis.	**L3**	**CO3**
13.	Given a dataset where different columns have varying levels of precision for numerical data (e.g., some values have two decimal places, others	**L3**	**CO3**

	have five), apply formatting to ensure uniform precision across all numerical columns. Justify why consistent precision is necessary.		
14.	Apply techniques to convert string representations of numerical data (e.g., '10,000', '5.5K') into numerical format for easy computation. Explain how this conversion enables accurate analysis and calculations.	**L3**	**CO3**
15.	Using a dataset with text-based categorical values (e.g., 'Male', 'male', 'M', 'F', 'Female', 'female'), apply techniques to standardize the categories to a uniform format. Explain why this step is important for further analysis.	**L3**	**CO3**
16.	Given a dataset where date fields are in different formats (e.g., `MM/DD/YYYY`, `DD-MM-YYYY`, `YYYY.MM.DD`), apply the necessary formatting to standardize all date fields to the `YYYY-MM-DD` format. Explain how this consistency improves data analysis.	**L3**	**CO3**
17.	Given a dataset with numerical features that have different ranges, apply min-max normalization to scale the values between 0 and 1. Show the steps and explain how this normalization improves model performance.	**L3**	**CO3**
18.	Using a dataset with both large and small magnitude features, apply Z-score normalization (standardization).	**L3**	**CO3**
19.	Apply log normalization to a dataset where the distribution of a feature is highly skewed. Compare the original distribution with the log-transformed one and explain how normalization affects skewed data.	**L3**	**CO3**
20.	Given a dataset with features measured on different scales (e.g., age in years, income in dollars), apply both min-max and Z-score normalization. Compare the results and discuss when each method would be more appropriate.	**L3**	**CO3**
21.	Given a dataset with a continuous variable such as age, apply equal-width binning to divide the data into 5 bins. Show the resulting bin intervals and explain how binning simplifies the analysis of continuous data.	**L3**	**CO3**
22.	Using a dataset with a skewed income distribution, apply equal-frequency binning to group the income variable into 4 bins, each containing an equal number of observations. Demonstrate the process and discuss why equal-frequency binning is useful in this context.	**L3**	**CO3**
23.	Given a numerical dataset with temperature values, apply custom binning to categorize the data into 'Low', 'Medium', and 'High' based on domain knowledge. Justify your binning strategy and explain how it improves interpretability.	**L3**	**CO3**
24.	Apply both equal-width and equal-frequency binning to a continuous feature in a dataset. Compare the results of the two binning methods and discuss when each method is more appropriate for analysis or model building.	**L3**	**CO3**
25.	Using a dataset with a categorical feature such as 'Gender' with values 'Male' and 'Female', apply **label encoding** in Python. Demonstrate how the categorical values are transformed into numerical values and explain why label encoding might not always be suitable for nominal data	**L3**	**CO3**
26.	Given a dataset with a categorical feature containing multiple categories (e.g., 'Country'), apply **one-hot encoding** in Python using `pandas.get_dummies()`. Show the transformed dataset and explain how one-hot encoding resolves the issue of imposing ordinal	**L3**	**CO3**

	relationships on nominal data.		
27.	Using a dataset with an ordinal categorical feature (e.g., 'Education Level' with values 'High School', 'Bachelor', 'Master', 'PhD'), apply **ordinal encoding** in Python by assigning appropriate numerical values based on order. Justify the assigned numerical values and explain why this method works well for ordinal data	**L3**	**CO3**
28.	Given a dataset with multiple categorical features, apply **frequency encoding** to a feature (e.g., 'Job Title'). Show how frequency encoding is applied using Python and discuss when frequency encoding is more appropriate than one-hot encoding.	**L3**	**CO3**
29.	Using a dataset with a categorical feature containing rare categories (e.g., 'Product Type'), apply **target encoding** to transform the categories based on the average target value. Demonstrate the steps in Python and explain how this method can improve model performance while avoiding overfitting.	**L3**	**CO3**
30.	Evaluate a scenario where a dataset contains outliers in numerical features. Discuss how different normalization techniques (min-max, Z-score, log transformation) can be applied in this context. Analyze which method would be most effective in preserving the integrity of the data and why.	**L4**	**CO3**
31.	Given a dataset where some features are already normalized while others are not, assess the potential consequences of using this dataset for model training without uniform normalization. Discuss how the lack of consistency in normalization can affect the results and interpretation of the model.	**L4**	**CO3**
32.	Analyze a dataset with a wide range of numerical features that exhibit different scales (e.g., height in centimeters and weight in kilograms). Compare the effects of min-max normalization and Z-score normalization on the dataset's distribution. Discuss how each method impacts the performance of a machine learning model trained on this dataset.	**L4**	**CO3**
33.	Given a dataset where a categorical variable represents different product types (e.g., 'Electronics', 'Clothing', 'Furniture'), analyze the impact of using one-hot encoding versus label encoding on a machine learning model. Discuss the scenarios in which each encoding technique may lead to better model performance and interpretability.	**L4**	**CO3**
34.	Consider a dataset with an ordinal categorical variable, such as 'Customer Satisfaction' rated from 'Poor', 'Average', to 'Excellent'. Analyze how different encoding strategies (e.g., ordinal encoding and target encoding) could influence the outcome of a regression model. Discuss the advantages and disadvantages of each approach in terms of maintaining the ordinal relationship.	**L4**	**CO3**
35.	Evaluate a dataset where a categorical variable has a high cardinality (e.g., 'City' names) and is used in a predictive model. Discuss the effectiveness of using target encoding compared to one-hot encoding in this context. Analyze the risks associated with target encoding and suggest strategies to mitigate these risks while retaining useful information.	**L4**	**CO3**
36.	Consider a scenario where two different data pre-processing pipelines are	**L4**	**CO3**

	applied to the same dataset: one focusing on scaling numerical features (using standardization) and the other on encoding categorical variables (using one-hot encoding).		
37.	Analyze a dataset that contains missing values, outliers, and mixed data types. Discuss how the choices made during data pre-processing (e.g., imputation methods, outlier removal techniques, and data type conversions) can affect the integrity of the dataset and the outcomes of subsequent analysis or machine learning models. Provide specific examples to support your analysis.	**L4**	**CO3**
38.	Analyze the impact of data binning on the distribution of a continuous variable, such as income. Compare the results of equal-width binning and equal-frequency binning in terms of information loss, interpretability, and model performance. Discuss scenarios where one method may be preferred over the other	**L4**	**CO3**
39.	Given a dataset with a highly skewed feature, evaluate how binning can be used to transform the feature into a more normal-like distribution. Discuss the implications of this transformation on the performance of machine learning models and the potential trade-offs involved in the binning process.	**L4**	**CO3**
40.	Consider a dataset where a categorical variable has been converted into bins (e.g., age groups: '0-18', '19-35', '36-50', '51+'). Analyze how this binning affects the overall analysis and interpretation of the data. Discuss the potential benefits and drawbacks of using binning for categorical variables versus continuous variables.	**L4**	**CO3**
41.	Evaluate the effects of using custom binning strategies based on domain knowledge compared to automated binning techniques (like k-means binning). Discuss how the choice of binning strategy can influence the insights gained from the data and the performance of predictive models.	**L4**	**CO3**
42.	Given a dataset that will be used for a classification task, analyze the trade-offs between retaining continuous variables versus binning them. Discuss how binning can affect model complexity, interpretability, and performance metrics, providing examples of when binning may enhance or hinder the analysis.	**L4**	**CO3**

UNIT –IV

Data Science in Business

Sr. No	Questions	Bloom's Level	CO Mapped
1.	How could predictive analytics in data science be applied to prevent potential outbreaks of infectious diseases? (Application of predictive models to real-world health scenarios)	**L3**	**CO4**
2.	In what ways could machine learning algorithms be utilized to improve patient diagnosis and treatment plans? (Exploring practical applications of machine learning in healthcare)	**L3**	**CO4**
3.	How would you apply data from wearable health devices to monitor and improve individual health outcomes? (Using data from health devices in a personalized healthcare context)	**L3**	**CO4**
4.	How might data science tools help identify risk factors for chronic diseases such as diabetes or heart disease? (Applying risk factor analysis techniques from data science to healthcare challenges)	**L3**	**CO4**
5.	How would you use AI-driven data models to optimize hospital resource allocation during an emergency situation? (Applying AI to manage healthcare resources efficiently)	**L3**	**CO4**
6.	How would you apply the concept of business problem identification when helping a company decide how to get started in data science?	**L3**	**CO4**
7.	Imagine you are consulting for a small company new to data science. How would you apply the CRISP-DM methodology to help them launch their first data science project?	**L3**	**CO4**
8.	Based on your knowledge of data infrastructure, how would you recommend a company begin setting up its data systems to support future data science initiatives?	**L3**	**CO4**
9.	Given a specific industry (e.g., healthcare or retail), how would you suggest a company prioritize the types of data to collect for its first data science project?	**L3**	**CO4**
10.	How would you apply the concept of a data-driven culture to help a company ensure successful adoption of data science practices across its teams?	**L3**	**CO4**
11.	Explain how data science techniques can be applied to predict customer behavior in e-commerce.	**L3**	**CO4**
12.	Demonstrate how data science can be used in the healthcare sector to predict patient outcomes based on historical data.	**L3**	**CO4**
13.	Apply clustering techniques to a real-world business scenario to segment customers into different groups based on purchasing behavior.	**L3**	**CO4**
14.	Illustrate the use of machine learning models in financial risk assessment by designing a predictive model for credit scoring.	**L3**	**CO4**
15.	Use data science methodologies to optimize a supply chain process, describing the steps and tools you would employ.	**L3**	**CO4**
16.	Apply the steps you would take to transition from a different career into data science. What resources or learning platforms would you use?	**L3**	**CO4**
17.	Given a specific problem in your field, how would you use data science skills to solve it? Which tools or techniques would be most appropriate?	**L3**	**CO4**
18.	Create a learning plan that outlines how you would gain the technical skills necessary to become a data scientist over the next year.	**L3**	**CO4**
19.	How would you implement machine learning models to analyze a dataset? What steps would you follow to ensure accuracy and reliability?	**L3**	**CO4**

20.	Given your current skill set, what projects could you undertake to build a portfolio showcasing your readiness to work as a data scientist?	**L3**	**CO4**
21.	Apply the steps you would take to transition from a different career into data science. What resources or learning platforms would you use?	**L3**	**CO4**
22.	Analyze the role of a data scientist in an e-commerce company. How does their work directly impact customer experience and sales growth?	**L3**	**CO4**
23.	Demonstrate how a data analyst's skill set can be applied to solving problems in the healthcare sector. What specific tasks would they perform?	**L3**	**CO4**
24.	Use your understanding of machine learning models to explain how a data engineer's responsibilities support data scientists in their analysis and modeling work.	**L3**	**CO4**
25.	Apply your knowledge of different career paths in data science to outline the steps a person would need to take to transition from a business analyst to a data scientist role.	**L3**	**CO4**
26.	Differentiate between the roles of a data scientist, data engineer, and machine learning engineer. How do their responsibilities and skill sets overlap or diverge?	**L4**	**CO4**
27.	**Examine** the potential career growth and opportunities within the field of data science. What factors influence the speed and direction of career progression?	**L4**	**CO4**
28.	**Analyze** the ethical challenges faced by data scientists working in sectors such as social media, healthcare, or finance. How might these challenges shape decision-making in data-driven projects?	**L4**	**CO4**
29.	**Compare** the required skills and educational backgrounds for a data scientist working in academia versus one working in a technology company. How does the application of their skills differ in each setting?	**L4**	**CO4**
30.	**Break down** the steps involved in transitioning from a traditional software engineering career to a data science role. What additional knowledge or experience is necessary to make this shift successful?	**L4**	**CO4**