DATA SCIENCE
(MCA2143)

Time Allotted : 2½ hrs **Full Marks : 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any 4 (four) from Group B to E, taking one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A

1. Answer any twelve: **12 × 1 = 12**

Choose the correct alternative for the following

(i) Why is data pre-processing important in data science?
(a) It ensures data is in a format suitable for analysis
(b) It helps in reducing computational complexity
(c) It improves the accuracy of machine learning models
(d) All of the above

(ii) Which of the following is NOT a typical step in the data wrangling process in data science?
(a) Data Cleaning (b) Data Exploration
(c) Data Visualization (d) Data Analysis.

(iii) What is the probability that a leap year has 53 Sundays?
(a) $\frac{1}{7}$ (b) $\frac{2}{7}$ (c) $\frac{3}{7}$ (d) None of these.

(iv) If X has a one parameter exponential distribution with pdf $f(x) = \lambda e^{-\lambda x}, x \geq 0$ Then variance is
(a) $\frac{1}{\lambda}$ (b) $\frac{1}{\lambda^2}$ (c) λ^2 (d) None of these.

(v) In which category does linear regression belong to?
(a) Neither supervised nor unsupervised learning
(b) Both supervised and unsupervised learning
(c) Unsupervised learning
(d) Supervised learning.

(vi) Which of the following is correct with respect to residuals?
(a) Positive residuals are above the line, negative residuals are below
(b) Positive residuals are below the line, negative residuals are above
(c) Positive residuals and negative residuals are below the line
(d) All of the mentioned.

(vii) K-Nearest Neighbours (KNN) is classified as what type of machine learning algorithm?
(a) Instance-based learning (b) Parametric learning
(c) Non-parametric learning (d) Model-based learning.

(viii) Which algorithm is best suited for a binary classification problem?
(a) K-nearest Neighbours (b) Decision Trees
(c) Random Forest (d) Linear Regression.

(ix) Which of the following is NOT a commonly used data visualization tool in data science?
(a) Plotly (b) Seaborn (c) TensorFlow (d) Matplotlib.

(x) What is true about Data Visualization?
(a) Data Visualization is used to communicate information clearly and efficiently to users by the usage of information graphics such as tables and charts.
(b) Data Visualization helps users in analysing a large amount of data in a simpler way.
(c) Data Visualization makes complex data more accessible, understandable, and usable.
(d) All of the above

Fill in the blanks with the correct word

(xi) For a random variable $E\{(X-2)^2\}=6, E\{(X-1)^2\}=10$. Then σ_x=_____

(xii) If a random variable X follows poisson distribution such that P(1)=P(2).Then Mean(X)=_____ .

(xiii) BeautifulSoup and Scrapy are popular Python libraries used for _____ in web scraping projects.

(xiv) Scatter plots are ideal for visualizing the _____ relationship between two continuous variables.

(xv) Data _____ involves the systematic gathering of information, ensuring its accuracy, completeness, and relevance for analysis and decision-making in data science projects.

Group - B

2. (a) When Data Cleaning is necessary in the dataset? What are the step we used for data cleaning? *[(CO1)(Identifying/LOCQ)]*
(b) How can we handle missing data if it is more in the dataset? *[(CO4)(Analyse/IOCQ)]*
(c) What do you mean by time series data? *[(CO4)(Clarifying/LOCQ)]*

6 + 4 + 2 = 12

3. (a) What are some common challenges faced when dealing with unstructured data in data science projects, and what strategies or techniques can be employed to effectively address these challenges? *[(CO1,2)(Understand/LOCQ)]*

(b) How does the choice of data cleaning techniques impact the reliability and accuracy of insights derived from a dataset during the data wrangling process? *[(CO1,2)(Remember/LOCQ)]*

(c) Suppose you have a dataset containing the ages of individuals from a certain population. You want to bin this data into age groups for analysis. The following are the age ranges and corresponding frequencies of individuals in each range:

Age Range	Frequency
0-10	25
11-20	45
21-30	60
31-40	55
41-50	40
51-60	30

Calculate the mean age for each binning range, rounded to the nearest integer. Then, find the overall mean age for this population dataset, also rounded to the nearest integer. *[(CO3)(Apply/IOCQ)]*

3 + 3 + 6 = 12

Group - C

4. (a) 3 balls are drawn at random from a bag containing 5 white and 3 black balls. Let. X be a random variable which denotes the number of white balls drawn. Find the distribution of X. *[(CO2)(Analyse/IOCQ)]]*

(b) A car hire firm has 2 cars which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson distribution with average number of demands per day 1.5. Evaluate the proportion of days on which some demand is refused. [Given $e^{-1.5} = 0.223$]. *[(CO5,2)(Apply/HOCQ)]*

5 + 7 = 12

5. (a) Compute the standard error of the mean and construct the sampling distribution of the mean for simple random samples of two families each from a population of 5 families which is given below : (suppose this is SRSWOR)

Family : A B C D E
Family Size : 4 3 2 5 7 *[(CO2) (Evaluate/HOCQ)]*

(b) Find the minimum number of times that a die has to be thrown such that the probability of 'no six' is less than $\frac{1}{2}$. *[(CO2, 5) (Apply/IOCQ)]*

6 + 6 = 12

Group - D

6. (a) How to determine k using the silhouette method? *[(CO6) (Analyse /IOCQ)]*

(b) Why is the odd value of "K" preferable in KNN algorithm? Why should we not use KNN algorithm for large datasets? *[(CO4)(Remember/LOCQ)]*

6 + 6 = 12

7. (a) What are density reachability and density connectivity? How does the epsilon value affect the DBSCAN Clustering Algorithm? *[(CO3)(Analyse/HOCQ)]*

(b) Explain with relevant example that how Euclidean Distance is measured in the K-Nearest Neighbors Algorithm. Explain the advantages and disadvantages of using K-NN algorithm. *[(CO3)(Understand/LOCQ)]*

6 + 6 = 12

Group - E

8. (a) Write a Python code snippet using Matplotlib to create a bubble chart representing the relationship between three variables: x-axis for age, y-axis for income, and bubble size for expenditure. Use appropriate labels and color coding to enhance readability. *[(CO3,4)(Apply/IOCQ)]*

(b) A researcher is analyzing the distribution of student grades in a class using a heat map. The grades are categorized into four groups: A (above 75%), B (60-75%), C (45-60%), and D (below 45%). If the heat map indicates that 20% of students received grade A, 30% received grade B, 35% received grade C, and 15% received grade D, evaluate the effectiveness of the visualization in conveying the distribution of grades. *[(CO4,5)(Evaluate/HOCQ)]*

6 + 6 = 12

9. (a) Define Visual Analytics. Why and when do we use Graph database? *[(CO6) (Remember/LOCQ)]*

(b) What is Z-Score method? Explain the application with a suitable example. *[(CO6) (Analyse/IOCQ)]*

(c) What type of data is best visualized with a heat map? *[(CO4) (Analyse/IOCQ)]*

(2 + 3) + (2 + 3) + 2 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	32.29	41.67	26.04