TOTAL PAGES: 2
20SSP60

PSG COLLEGE OF ARTS & SCIENCE
(AUTONOMOUS)

MSc(SS) DEGREE EXAMINATION DECEMBER 2025
(Ninth Semester)

Branch – **SOFTWARE SYSTEMS (five years Integrated)**

DATA SCIENCE

Time: Three Hours Maximum: 50 Marks

SECTION-A (5 Marks)

Answer **ALL** questions
ALL questions carry **EQUAL** marks (5 x 1 = 5)

1. What is the main purpose of data cleaning?
   (i) To visualize data (ii) To remove or correct inaccurate data
   (iii) To store data in a database (iv) To encrypt dat

2. What is the main advantage of using MapReduce for data processing at scale?
   (i) It requires a central server for computation
   (ii) It allows sequential execution of all map tasks before reduce tasks
   (iii) It enables distributed processing of large datasets across clusters
   (iv) It is limited to structured data only

3. Which of the following best describes a spline in regression analysis?
   (i) A polynomial used for classifying data into categories
   (ii) A smooth piecewise polynomial function used to fit data flexibly
   (iii) A linear model that fits only straight lines
   (iv) A clustering technique for large datasets

4. Which of the following is a key advantage of using the Singular value Decomposition (SVD) in data analysis?
   (i) It automatically removes all missing values
   (ii) It helps in dimensionality reduction and identifying rank-deficiency
   (iii) It converts categorical data to numerical data
   (iv) It guarantees linear independence of features

5. Descriptive statistics and visualizations are primarily used to:
   (i) Build predictive models for future data
   (ii) Summarize, explore, and communicate patterns in data
   (iii) Perform QR decomposition
   (iv) Generate synthetic datasets only

SECTION - B (15 Marks)

Answer **ALL** Questions
ALL Questions Carry **EQUAL** Marks (5 x 3 = 15)

6 a. Evaluate a given raw dataset (e.g., customer transactions) and identify the key steps of data wrangling and cleaning that would improve its quality. Justify why each step is necessary for accurate analysis.

OR

b. Describe different measures of central tendency and variability.

7 a. Given a large dataset and multiple processors, explain how a parallel database system can be applied to improve query performance. Illustrate with an example of data partitioning or parallel query execution.

OR

b. What is MapReduce?Explain the working principle of MapReduce with a neat diagram and discuss its advantages in handling big data.

Cont...

8 a Design a small experiment to collect data, perform univariate linear regression, and visualize the results.

OR

b What are splines in regression analysis?Discuss how splines improve model flexibility compared to simple linear regression, with an example.

9 a State and explain the Gauss–Markov Theorem. Discuss its significance in linear regression.

OR

b Describe the QR decomposition method.Explain how QR decomposition and Gram–Schmidt orthogonalization are used in solving least squares problems.

10 a. Explain the process of community detection with a suitable case study.

OR

b. Describe how collaborative networks are analyzed using graph analytics techniques.

SECTION -C (30 Marks)

Answer **ALL** questions

ALL questions carry **EQUAL** Marks (5 x 6 = 30)

11 a Discuss the importance of mathematics in understanding data with examples supporting data interpretation and modeling.

OR

b Explain the role of data visualization in descriptive statistics and describe various charts and plots used for data exploration.

12 a Explain the concept of data partitioning in parallel databases and its impact on query efficiency and load balancing.

OR

b Explain how parallel query execution strategies differ in shared-nothing and shared-disk database architectures.

13 a Elucidate how rank deficiency affects regression solutions and discuss methods to handle it in linear models.

OR

b Describe the role of subspaces and orthogonal projections in solving least squares regression problems.

14 a Explain how estimable functions are identified in linear models and their role in hypothesis testing.

OR

b Describe how QR decomposition and Gram-Schmidt orthogonalization are applied in solving linear regression problems.

15 a Explain how graph analytics is applied to analyze large-scale networks and extract meaningful insights.

OR

b Discuss the techniques for recursive queries in graph databases and their significance in semantic web applications.

Z-Z-Z END