Registration Number:

Date & session:

ST JOSEPH'S UNIVERSITY, BENGALURU -27
M.SC (BIG DATA ANALYTICS) – II SEMESTER
SEMESTER EXAMINATION: APRIL 2024
(Examination conducted in May / June 2024)
BDADE2621 – MULTIVARIATE STATISTICS
(For current batch students only)

Time: 2 Hours **Max Marks: 50**

This paper contains TWO printed pages and THREE parts.

PART-A

Answer ALL the questions **2 X 5 = 10**

1. Provide two real-life examples where cluster analysis is utilized.
2. What is survival analysis?
3. Calculate the Jaccard and cosine similarity between two sets A and B given the following information: Set A = {5, 3, 2, 1} and Set B = {1, 2, 3, 4}
4. What a short note on R Squared and Adjusted R- Squared.
5. State the assumptions of multiple linear regression models.

PART-B

Answer Any FIVE questions **4 X 5 = 20**

6. List the limitations of factor analysis?
7. Distinguish between univariate, bivariate and multivariate statistics.
8. Describe the concept of multicollinearity and its implications in multivariate regression analysis. Mention two methodologies to diagnose violations of the assumptions underlying a multiple linear regression model.
9. Demonstrate K Means clustering with example
10. Write a short note on the revolution of LLMs.
11. How Multivariate techniques helps in improving the social media analytics.
12. A researcher conducted a study to investigate the relationship between three variables: screen time (X1), ability to focus in minutes (X2), and income (Y). The following covariance matrix was obtained:

	X1	X2	Y
X1	16	8.5	-14
X2	8.5	9	-6
Y	-14	-6	25

a) Calculate the correlation matrix for the variables.

b) Interpret the correlation coefficient between income and screen time.

PART-C

Answer Any TWO questions **10 X 2 = 20**

13. Describe the following with respect Principal Component Analysis (PCA).
    a) Benefits of PCA [5]
    b) Eigen values and Eigen vectors in PCA. [3]
    c) Covariance Matrix [2]
14. State and explain Hierarchical Clustering Algorithm. How do you interpret the Dendrogram?
15. A tech entrepreneur wishes to develop the next revolutionary smartphone app in 2024. Formulate this as a multiple regression problem and propose a solution keeping the following in mind:
    a) What is the dependent variable, what could be the independent predictors? [4]
    b) What is collinearity? Could there be a presence of collinearity? [3]
    c) Which one would you use ‘R squared’, or ‘adjusted R squared? Why? [3]