Register Number:

DATE:

ST. JOSEPH’S COLLEGE (AUTONOMOUS), BENGALURU-27
M.Sc. BIG DATA ANALYTICS – II SEMESTER
SEMESTER EXAMINATION: APRIL 2019
BDADE 2518 - Multivariate Statistics

TIME 2.5 HOURS **MAXIMUM MARKS 70**

This Question Paper Contains TWO Printed Paper And ONE Part

Answer any 5 questions. Each question carries 14 marks

1a Explain the idea of analysis of variance (ANOVA) |6|
1b Using any example, write down the null and alternate hypotheses of the problem |4|
1c How do you compute the F ratio? What’s the underlying rationale |4|

2. Let X be your expected mark in this exam. Let Y be the number of hours that you studied for this exam. Create a dummy X-Y data set for 5 students and then:

- Compute the correlation coefficient between X and Y |4|
- Write down the regression equation of Y (dependent variable) on X |4|
- Explain the idea of least squares with a sketch |6|

3. A billionaire wishes to produce the blockbuster Malayalam film of 2020. Formulate this as a multiple regression problem and propose a solution keeping the following in mind:

- What is the dependent variable, what could be the independent predictors?|5|
- What is collinearity? Could there be presence of collinearity? |3|
- Would you use ‘R squared’, or ‘adjusted R squared?’ Which one? Why? How? |6|

4a What is the rationale of principal component analysis? |5|
4b Sketch (as a flow chart) the different steps involved in PCA |5|
4c Mention two applications where PCA can make a big difference |4|

5a What are the different ways of clustering? |4|

5b Give two real-life examples (from sport of business) where we use cluster analysis |4|

5c Pick any one clustering algorithm and draw a flow chart to implement it |6|

6 A telecom company is worried about customer churn and seeks your help. Discuss how you would solve this problem.

- What is the meaning of customer churn? |4|
- Make a list of 4-5 predictive variables |4|
- What is logistic regression? How would you use it to solve this problem? |6|

7 Write short notes on any two of the following: |7+7|

- Eigen values and eigen vectors
- Multivariate techniques in social media analytics
- Python vs R? What’s your choice, and why?