Register No:

Date:

ST. JOSEPH'S COLLEGE (AUTONOMOUS), BANGALORE-27
M.SC. BIG DATA ANALYTICS - II SEMESTER
SEMESTER EXAMINATION: APRIL 2018
BDADE 2516: MULTIVARIATE STATISTICS

TIME: 2 ½ HRS **MAX MARKS 70**

This Question Paper Contains ONE Printed Pages

Answer as many questions as possible but maximum 70 marks

1a Explain the idea of analysis of variance using the example of 1-way ANOVA |6|
1b Give an example where you might need to use a 2-way ANOVA |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. Discuss how you can convert the bivariate problem of Question 2 into a multivariate problem. Specifically highlight the following points (don't write more than one page in all)

- New independent variables you might add |4|
- The Probable presence of collinearity |4|
- Using R squared, or adjusted R squared? Which one? Why? |6|

4a Describe (in no more than 5 sentences) the benefits 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 is the underlying principle of clustering? |4|
5b Give two real-life examples (from sport of business) where cluster analysis helps |4|
5c Give a step-by-step description of how to do k-means clustering |6|

6 A bank has a tricky decision to make. Should it offer a credit card to a customer with a seemingly modest income?

- What is logistic regression? Why should you use it to solve this problem? |6|
- List out 10-12 possibly predictive variables? |4|
- Sketch (as a flow chart) your options of stepwise regression |4|

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

- Eigen values and eigen vectors
- Multivariate techniques in HR analytics
- Why the correlation coefficient is better than the covariance