Registration Number:

Date & session:

ST JOSEPH'S UNIVERSITY, BENGALURU -27
M.Sc (STATISTICS) – 2nd SEMESTER
SEMESTER EXAMINATION: APRIL 2024
(Examination conducted in May / June 2024)
ST8521 – INTRODUCTION TO DATA SCIENCE
(For current batch students only)

Time: 1 Hours **Max Marks: 25**

This paper contains TWO printed pages and ONE part

PART-A

Answer any FIVE from the following question with respect to the case studies stated below. **5x5=25**

Case Study 1: Suppose a given dataset contains the information about bank loans granted to various customers. The dataset includes details such as loan amount(float), interest rate(float), customer credit score(int), loan term(float), customer demographics (str) and loan status (approved or denied) (Binary).

Case Study 2: A survey was conducted on heart failure patients, a critical area of study in clinical research. The dataset contains various features related to patient health, lifestyle, and medical history. However, like any real-world dataset, it's not pristine; missing values lurk within its rows and columns and maybe some exceptional datapoints too.

Case Study 3: A reputable hospital aims to enhance patient satisfaction by understanding the factors that influence it. The hospital collects data from 46 patients, focusing on four key variables:

Satisfaction: A continuous variable representing the degree of satisfaction with the quality of care. Higher values indicate greater satisfaction.
Age: The age of each patient in years.
Severity: A continuous variable indicating the severity of the patient's medical condition. Higher values correspond to more severe cases.
Stress: The patient's self-reported level of stress. Higher values reflect higher stress levels.

[Answer Q1, Q2 and Q3 with reference to case study 1]

1. a. Suppose you are using excel software to analyse the data, Identify the correct tool to visually compare loan amount and credit score of the customers and outline the steps for obtaining the same.
   b. With the data being available, explain any 3 different functions that one will use to do the preliminary analysis? (2+3)

2. Explain to your friend, who is a beginner in R, about the different conditional statements and loops in brief for advanced data analysis within the R software.

3. Identify the relevant clusters and elaborate on any two methods of finding clusters.

[Answer Q4 and Q5 with reference to case study 2]

4. Discuss different types of missing values imputing techniques with an example.

5. For the above case study, give a detailed write up on one graphical and analytical method to detect outliers.

[Answer Q6 and Q7 with reference to case study 3]

6. Your friend, who works in the hospital's management department, wants to analyse the data using a multiple linear regression model. Can you provide a detailed explanation of the model to help them understand it clearly?
   Ans: A detailed note on multiple linear regression to be discussed

7. To enhance model efficiency, you opt to employ Forward and backward stepwise regression model selection technique. Elaborate on the same.

****