Question Paper

Exam Date & Time: 25-Jun-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VI Semester Makeup Examinations - June 2024

INTRODUCTION TO DATA SCIENCE [CRA 4060]

Marks: 50 **Duration: 180 mins.**

Descriptive

Answer all the questions. Section Duration: 180 mins

1A) Using the **dplyr** package in R, perform the following tasks on the given dataset**students_data**: (5)

i. Filter the dataset to include only students with a score above 80 in the "Math" subject.

ii. Group the filtered dataset by the "Gender" column.

iii. Calculate the average score in the "Science" subject for each gender group.

iv. Rename the columns to "Gender" and "Average_Science_Score".

v. Write the resulting dataset to a new CSV file named "gender_avg_science.csv".

1B) Compare the height distributions of two species of trees, A and B, in a forest reserve by constructing two side-by-side box plots. The heights of sampled trees for each species are provided as follows: Tree A (25, 30, 32, 35, 36, 38, 40, 41, 42, 45) and Tree B (20, 22, 25, 28, 30, 32, 34, 36, 38, 40). Alongside the box plots, calculate and display the minimum height, lower quartile (Q1), median, upper quartile (Q3), and maximum height for each species. (3)

1C) Discuss the role of exploratory plots in data analysis, highlighting their importance in uncovering patterns, trends, and relationships within datasets. Provide two examples of exploratory plots commonly utilized in data analysis and describe how each plot type aids analysts in gaining insights into the data. (2)

2A) How would you perform hierarchical clustering on a dataset containing ***gene expression*** data using R? Detail the steps involved, including loading necessary packages, data preprocessing, applying hierarchical clustering with complete linkage, and visualizing the resulting dendrogram. Discuss the interpretation of the dendrogram and its utility in identifying clusters of genes with similar expression patterns, highlighting its significance in biological research and data analysis. (5)

2B) Using R, perform k-means clustering on a dataset containing customer transaction data. After loading the necessary packages and preprocessing the data if needed, apply k-means clustering with the appropriate number of clusters. Print the cluster centers and create a visualization to illustrate the clustering results. Discuss how these clusters can be utilized by the retail company to improve marketing strategies and enhance customer satisfaction. (3)

2C) Discuss the advantages of using the lattice system for creating plots compared to other plotting systems in R, such as base graphics or ggplot2. (2)

3A) Illustrate how to enhance a plot by incorporating color transparency and labels using the ggplot2 package in R. Provide R code demonstrating these enhancements along with explanations for each step. (5)

3B)	Discuss the challenges in determining an ideal data set during data analysis.	(3)
3C)	Identify functions from the cacher package in R which performs the following tasks: i. Clones a cached object identified by the specified ID. ii. Displays the files stored in the cache directory. iii. Displays the structure of the code in the loaded R script file iv. Loads an R script file	(2)
4A)	Discuss the pros and cons ofreproducibility in data analysis and research	(5)
4B)	Discuss the necessities for reproducible research.	(3)
4C)	Elaborate on the limitations encountered when employing Sweave as a tool for literate programming in R.	(2)
5A)	With the help of a diagram, discuss the research pipeline in reproducible research.	(5)
5B)	Discuss the various aspects that must be kept track of in the software environment in a reproducible research checklist.	(3)
5C)	Compare and contrast Replication and Reproducibility in reproducible research.	(2)

-----End-----