Question Paper

Exam Date & Time: 01-Jun-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, JUNE 2023

INTRODUCTION TO DATA SCIENCE [CRA 4060]

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

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed

1) Using R code, derive a single cluster by employing hierarchical clustering approach from the following distance matrix. (5)

A)

Dist	A	B	C	D	E	F
A	0.00	0.71	5.66	3.61	4.24	3.20
B	0.71	0.00	4.95	2.92	3.54	2.50
C	5.66	4.95	0.00	2.24	1.41	2.50
D	3.61	2.92	2.24	0.00	1.00	0.50
E	4.24	3.54	1.41	1.00	0.00	1.12
F	3.20	2.50	2.50	0.50	1.12	0.00

B) Consider the dataset *msleep* (mammals sleep) with 83 rows and 11 variables, which contains the sleep times and weights for a set of mammals. Description is given in the following table. (3)

column name	Description
name	common name
genus	taxonomic rank
vore	carnivore, omnivore or herbivore?
order	taxonomic rank
conservation	the conservation status of the mammal
sleep_total	total amount of sleep, in hours
sleep_rem	rem sleep, in hours
sleep_cycle	length of sleep cycle, in hours

awake	amount of time spent awake, in hours
brainwt	brain weight in kilograms
bodywt	body weight in kilograms

Write R code to perform the following operations on the given dataset using *dplyr* verbs.

i. Load the dataset, list the mammals who sleeps more than 16 hours and arrange them in the order of total hours they sleep.

ii. Create a new column called *rem_proportion*(ratio of rem sleep to total amount of sleep) and create summary statistics(mean) for the column sleep_total.

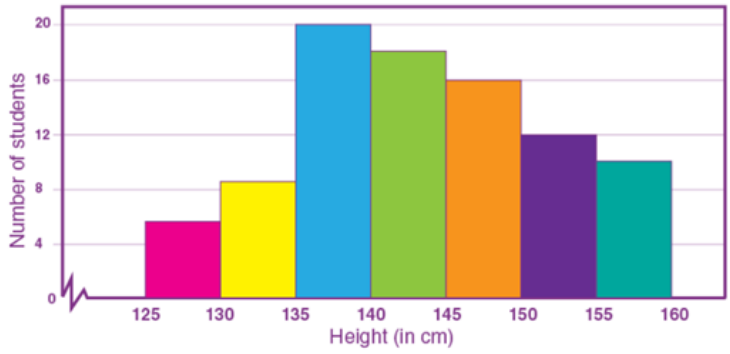
C) Using a R code example, illustrate how to add labelstoaplot by employing ggplot2y. (2)

2) Consider the set of multivariate variables N_1 ..... Nn where $N_1 = (N_{11}.....N_{1m})$, explain how to find the new set of variables that are uncorrelated and how to find one best matrix which depicts the original data with fewer variables. (5)

A)

B) With the help of R code, describe the process of building heatmaps from K - means solutions. (3)

C) The following histogram shows the height of students within a class (2)



i. How many students have a height of less than 140 cm?

ii. How many students have a height greater than 140 cm but less than 155 cm?

3) Describe different lattice functions available with suitable R commands and examples. (5)

A)

B) Summarize the significance of Markdown and RMarkdown in data analysis. (3)

C) Identify key data analysis files that are typically produced while a supply chain data analysis is performed. (2)

4) Explain the significance of cacher package in R. List and explain the major functions used in R to implement caching computations. (5)

A)

B) Outline the effectiveness of literate programming in data analysis and analyze the different packages used for this purpose? (3)

C) Why is it important to replicate studies in data science, and how does it contribute to the overall reliability and credibility of research findings? (2)

5) Discuss the best practices that researchers should follow to ensure reproducibility in their research, and what are some common pitfalls or bad practices that they should avoid? (5)

A)

B) Suppose a pharmaceutical company has developed a new drug for treating a specific disease, and they want to know whether it is effective. Discuss how evidence based analysis can help the researchers in providing a solution for the given problem. (3)

C) Elaborate the process of generating HTML document using knitr package. (2)

-----End-----