Question Paper

Exam Date & Time: 08-May-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH END SEMESTER EXAMINATIONS, MAY 2024

INTRODUCTION TO DATA SCIENCE [CRA 4060]

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

A

Answer all the questions. Section Duration: 180 mins

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

1) A) Consider the dataset **customer_data** which contains information about customers and their purchases. Use the **dplyr** package in R to perform the following tasks: First, filter the dataset to include only customers who have made purchases in the past month. Then, group the filtered dataset by the "Age_Group" column, which categorizes customers into different age brackets. Calculate the average purchase amount for each age group. Next, identify the top 3 age groups with the highest average purchase amounts and sort them in descending order based on their average purchase amounts. Finally, export the resulting dataset to a new CSV file named "top_age_groups.csv". Ensure to include appropriate column names and maintain consistency with the original dataset structure. (5)

B) Construct a complete box plot using the provided data on a class of students' exam scores. The data reveals the following statistical information: the largest recorded score is 92, the lower quartile is 73, the median score is 81, the range of scores spans 21 points, and the inter-quartile range is 11. (3)

C) Using R, create a histogram to visualize the distribution of ages in a dataset containing the following ages: 25, 28, 30, 32, 35, 35, 38, 40, 42, 45, 50, 55, 60, 65. Ensure that the histogram bins are appropriately chosen to represent the age distribution clearly and accurately. Additionally, label the horizontal axis as "Age" and the vertical axis as "Frequency". Write the R code to generate this histogram. (2)

2) A) Using R, create a scatter plot to visualize the relationship between two variables in a dataset, with each point colored according to a third categorical variable. Perform the following tasks: (5)

a. Load the necessary R packages and import the dataset.

b. Create a scatter plot with one variable on the x-axis, another variable on the y-axis, and points colored according to the third categorical variable.

c. Customize the plot by adding appropriate labels to the axes and a title.

d. Discuss the importance of incorporating color into data visualization and how it enhances the interpretation of relationships between variables in the plot.

e. Explain any potential limitations or considerations when using color in data visualization, and suggest strategies to mitigate them.

B) Using R code, discuss how heatmaps effectively represent dense data sets, such as gene expression data, by using color gradients to encode information. (3)

C) Using the lattice system in R, create a lattice plot to visualize the relationship between two variables, X and Y, in a dataset. Perform the following tasks:
a. Load the necessary R packages and import the dataset.
b. Create a lattice plot with Y on the vertical axis and X on the horizontal axis, using points to represent the data. (2)

3) Demonstrate how to customize the appearance of a bar plot created using ggplot2 in R. Include examples of adjusting the color palette, adding annotations, and modifying axis labels, each with accompanying R code. (5)

A)

B) Discuss the necessity of data cleansing prior to implementing any algorithms on it. (3)

C) Write a R markdown document to create a simple report to analyze the iris dataset which consists of Sepal length, Sepal width, Petal length, Petal width as attributes and Species as class label. (2)

4) Propose a methodology for analyzing the association between short-term changes in air pollution and short-term changes in a population health outcome, particularly focusing on time series data. (5)

A)

B) Using knitr in R Markdown create a report on the impact of air pollution on respiratory health outcomes. The dataset consists of multiple variables, including: (3)

Date: The date of measurement

Pollution Level: Measurements of air pollution indicators (e.g., PM2.5, NO2, etc.)

Respiratory Health Outcome: Various respiratory health metrics (e.g., asthma admissions, COPD mortality, etc.)

Other Covariates: Additional variables such as temperature, humidity, and socio-demographic factors that may influence respiratory health outcomes.

C) (2)

```
def add_numbers_but_between_100_and_10000(x, y):
  """Add two numbers and return the result."""
                result = x + y
  return result
def greet(name):
                """Greet the user with a personalized message."""
  message = f"Hello, {name}! Welcome to our program."
  print(message)
# Call the functions
sum_result = add_numbers(5, 3)
print("The sum is:", sum_result)
greet("Alice")
```

Which of the following lines of code violates the coding standard in R. Justify your answer

5) How can the hierarchy of information be structured to ensure reproducibility in a research paper? (5)

A)

B)	Discuss the steps to critically evaluate a research process or findings in reproducible research.	(3)
C)	Elaborate on the usage of version control systems in reproducible research.	(2)

-----End-----