Question Paper

Exam Date & Time: 10-Jul-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

SIXTH SEMESTER B.TECH MAKEUP EXAMINATIONS, JULY 2023

INTRODUCTION TO DATA SCIENCE [CRA 4060]

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

Descriptive

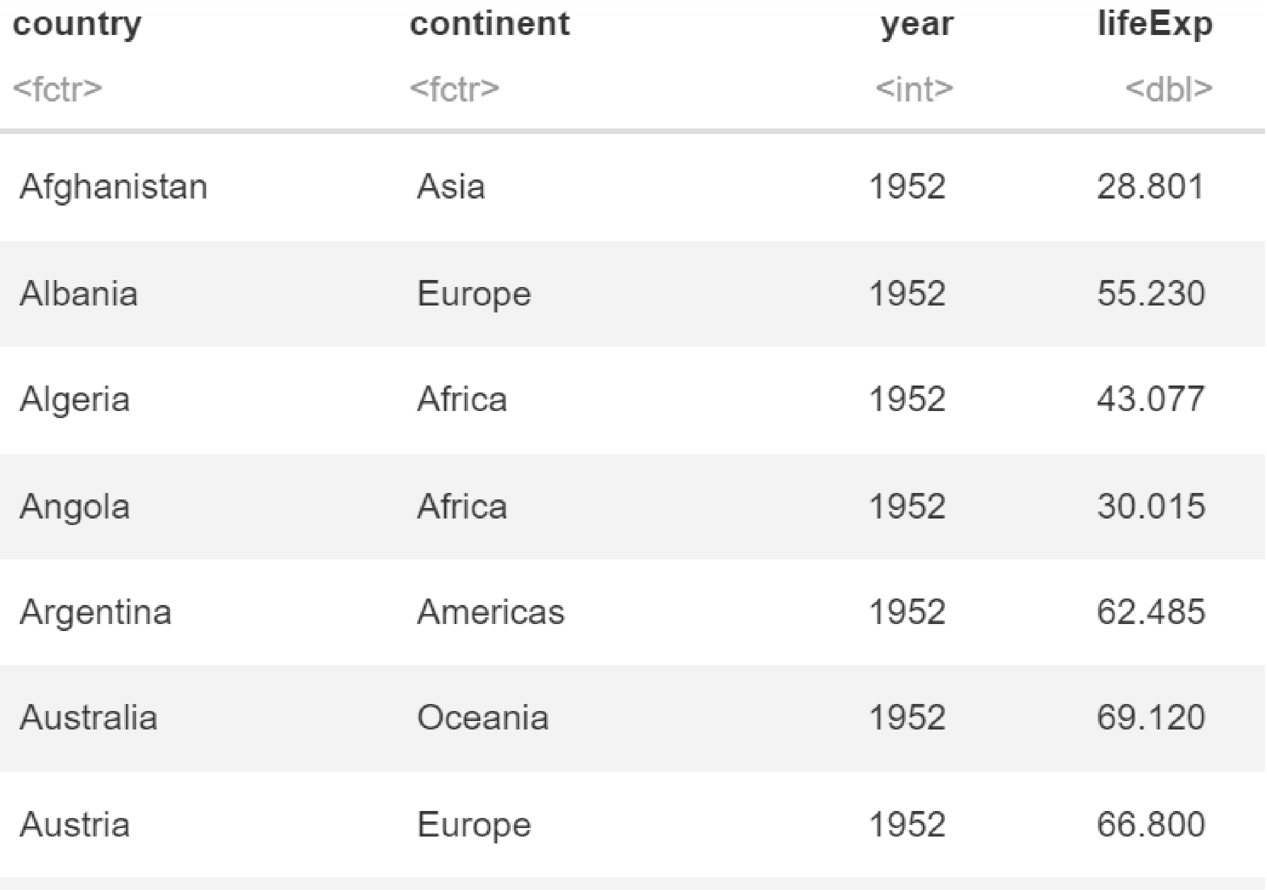
Answer all the questions.

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

1) Discuss the metrics used to find similarity between the data objects with the help of a real time application. (5)

A)

B) Consider the gapminder dataset given below and perform the following operations using dplyr verbs. (3)

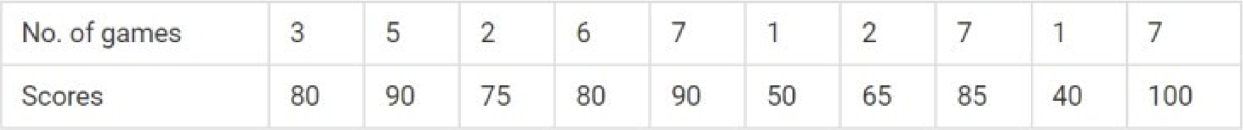


country <fctr>	continent <fctr>	year <int>	lifeExp <dbl>
Afghanistan	Asia	1952	28.801
Albania	Europe	1952	55.230
Algeria	Africa	1952	43.077
Angola	Africa	1952	30.015
Argentina	Americas	1952	62.485
Australia	Oceania	1952	69.120
Austria	Europe	1952	66.800

Write R code to perform the following operations on the given dataset using dplyr verbs.
i. Obtain the average life expectancy for each continent for each year.

ii. For a specific year (say 1952), show a graph to compare each of the countries with the average life expectancy of their continent.

C) Draw a scatter plot for the given data that shows the number of games played and scores obtained in each instance. (2)



No. of games	3	5	2	6	7	1	2	7	1	7
Scores	80	90	75	80	90	50	65	85	40	100

2) Using R code, illustrate the process of defining the cluster using K - means algorithm. (5)

A)

B) Consider a situation where the rows of a matrix represent observations of some sort and the columns of the matrix represent features or variables. Discuss how Singular Value Decomposition can be used to solve this problem? (3)

C) Discuss the purpose of major component in Principle Component Analysis for dimensionality reduction. (2)

3) Illustrate the components involved in the research pipeline from the perspective of author and reader. (5)

A)

B) Using a R code example, illustrate how to add colour transparency. (3)

C) With the help of R code, Illustarte the functionality of brewer.pal() and colorRampPalette(). (2)

4) What is the significance of literate programming in ensuring reproducibility in research? Suggest any two commonly used literate programming packages. (5)

A)

B) Discuss in brief the do's and don't of reproducibile research. (3)

C) Write the coding standards followed in R. (2)

5) Elaborate the given steps involved in data analysis with appropriate example, figures and pseudocode. (5)
a. Exploratory data analysis
A) b. Statistical prediction/modelling.
c. Create reproducible code.

B) Explain the significance of RMarkdown and knitr package in data analysis. (3)

C) Differentiate between Replication and Reproducibility. (2)

-----End-----