GALGOTIAS UNIVERSITY

ADMISSION NUMBER

School of Computing Science and Engineering

Master of Science in Computer Science
Semester End Examination - Jun 2024

Duration : 180 Minutes
Max Marks : 100

Sem II - E1PY205B - Introduction to Data Analytics

General Instructions
Answer to the specific question asked
Draw neat, labelled diagrams wherever necessary
Approved data hand books are allowed subject to verification by the Invigilator

1)	How do you import the Matplotlib library and rename it as plt for convenience?	K1(2)
2)	Discuss how the evolution of analytic scalability contribute to organizations' competitive advantage.	K2(4)
3)	Identify and discuss three applications of data analytics across different industries. Explain how data analytics has transformed processes and decision-making in these domains.	K2(6)
4)	Elaborate the process of creating a box plot using Seaborn. Consider the following dataset representing the monthly sales (in thousands of dollars) of a company over a year: Month Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Sales 50 55 60 65 70 75 80 85 90 95 100 105 Create a line plot using Seaborn to visualize the trend of sales over the year.	K3(9)
5)	Interpret the significance of pair plots in exploratory data analysis using Seaborn. Consider the following dataset representing the height and weight of individuals: Height (in inches) 65 67 68 70 72 Weight (in pounds) 140 150 155 160 170 Create a pair plot using Seaborn to visualize the relationship between height and weight.	K3(9)
6)	Given a dataset containing missing values, perform data cleaning using pandas. Discuss three common techniques for handling missing values and apply each technique to the dataset. Provide code snippets and explanations for each technique.	K5(10)
7)	Outline the significance of variance and standard deviation in analyzing the spread or variability of data points in real-world datasets. A company recorded the daily sales (in dollars) for a week: 1200, 1500, 1400, 1600, 1800, 1700, 2000. Calculate the	K4(12)

variance and standard deviation of the sales. Provide Python code for it and interpret the results to understand the data's variability.

8) A company has stored their employee data in a CSV file named "employee_data.csv". The file contains information about employees, including employee ID, name, age, gender, department, and salary. Perform the following tasks using Pandas: (a)Read the data from the CSV file into a DataFrame. (b)Check for any missing values in the dataset and handle them appropriately. (c)Transform the data to create a new column named "monthly_salary" which represents the monthly salary of each employee (salary divided by 12). (d)Normalize the "monthly_salary" column to ensure all values fall within the range [0, 1]. (e)Write the cleaned and transformed data back to a new CSV file named "cleaned_employee_data.csv". Provide code snippets and summary of your findings, including insights gained from the analysis. K5(15)

9) A research institute has stored their experiment data in a CSV file named "experiment_data.csv". The file contains information about experiments conducted, including experiment ID, researcher name, date, temperature, and success status (True or False). Perform the following tasks using Pandas: (a)Read the data from the CSV file into a DataFrame. (b)Check for any missing values in the dataset and handle them appropriately. (c)Transform the data to create a new column named "month" which represents the month of the experiment conducted. (d)Normalize the temperature column to ensure all values fall within the range [0, 1]. (e)Write the cleaned and transformed data back to a new CSV file named "cleaned_experiment_data.csv". Provide code snippets and summary of your findings, including insights gained from the analysis. K5(15)

10) Discuss the significance of conducting hands-on exercises using Python for statistical analysis. A company wants to analyze whether there is a significant difference in the average response time (in seconds) between two different customer service methods, Method A and Method B. They collected response time data from a sample of 30 customer interactions for each method. Conduct a two-sample t-test to determine if there is a significant difference between the average response times of the two methods at a 5% significance level. Provide the necessary Python code to perform the analysis. K6(18)