JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

Make-up Examination-Nov-2025

COURSE CODE (CREDITS): 22M11CI112 (3) MAX. MARKS: 25

COURSE NAME: INTRODUCTION TO DATA SCIENCE

COURSE INSTRUCTORS: Dr Nancy Singla MAX. TIME: 1 Hour 30 Minutes

Note: *(a) All questions are compulsory.*
(b) The candidate is allowed to make suitable numeric assumptions wherever required for solving problems and use of calculator is allowed.

Q. No.	Question	CO	Marks
Q1	You are given a CSV file containing monthly sales data from multiple regional offices. You notice the following issues: • Some rows have missing values in the `Revenue` and `Salesperson` columns. • The `Date` column is in multiple formats (e.g., `"2025/03/01"` and `"03-01-2025"`). • The `Region` column has inconsistent capitalization (e.g., `"north"`, `"North"`, `"NORTH"`). a) Describe step-by-step how you would clean and standardize this dataset. b) What methods would you use to handle missing data appropriately? c) How would you detect and treat outliers in the `Revenue` column? d) Provide one Python or pandas function that could help with each issue.	CO3	[2*4=8]
Q2	You are developing a news article classifier that must categorize articles into topics like Politics, Sports, Technology, etc. The text dataset is large (10 million records). Which method from stemming or lemmatization would you choose and why? How does your choice impact accuracy and computational time?	CO3	[2]
Q3	An environmental agency tracks pollution levels with: • A time series chart showing daily AQI (Air Quality Index) over six months. • A heatmap of hourly AQI levels across weekdays. The time series indicates gradual AQI improvement after a ban on certain fuels, while the heatmap shows consistent peaks during morning and evening traffic hours. (a) Explain what each visualization contributes to the overall analysis. (b) How can these findings guide city traffic or pollution control policies?	CO4	[3+2=5]

35

<table>
<tr><td>Q4</td><td>A factory produces LED bulbs, and the probability that any bulb is defective = 0.05. A quality inspector randomly selects 10 bulbs.
(a) Define the random variable and its distribution parameters.
(b) Find the probability that exactly 2 bulbs are defective.
(c) Find the probability that at most 1 bulb is defective.</td><td>CO5</td><td>[1+2+2=5]</td></tr>
<tr><td>Q5</td><td>A marketing manager wants to predict sales (Y, in ₹'000) based on advertising expenditure (X, in ₹'000).
The data for 5 weeks is given below:
<table><tr><th>Week</th><th>Advertising (X)</th><th>Sales (Y)</th></tr><tr><td>1</td><td>2</td><td>4</td></tr><tr><td>2</td><td>4</td><td>6</td></tr><tr><td>3</td><td>6</td><td>7</td></tr><tr><td>4</td><td>8</td><td>9</td></tr><tr><td>5</td><td>10</td><td>11</td></tr></table>(a) Compute the regression equation of Y on X using the least squares method.
(b) Estimate the sales when the advertising spend is ₹7,000.
(c) Interpret the slope of the regression line.</td><td>CO6</td><td>[2+2+1=5]</td></tr>
</table>