


VIT

Vellore Institute of Technology
(Granted to be University under section 3 of the UGC Act, 1956)

Reg. No: _____

Final Assessment Test – November 2025

Course: BCSE331L

- Exploratory Data Analysis

Class NBR(s): 1620/1631

Time: Three Hours

Slot: C2

Max. Marks: 100

- KEEPING MOBILE PHONE/ANY ELECTRONIC GADGETS, EVEN IN 'OFF' POSITION IS TREATED AS EXAM MALPRACTICE
- DON'T WRITE ANYTHING ON THE QUESTION PAPER

COs	CO Statements
CO1	Handle missing data in the real world data sets by choosing appropriate methods.
CO2	Summarize the data using basic statistics. Visualize the data using basic graphs and plots.
CO3	Identify the outliers if any in the data set.
CO4	Choose appropriate feature selection and dimensionality reduction.
CO5	Apply Techniques for handling multi-dimensional data.

BL – Blooms Taxonomy Level (1 – Remember, 2 – Understand, 3 – Apply, 4 – Analyse, 5 – Evaluate, 6 – Create)

Answer ALL Questions

(10 X 10 = 100 Marks)

- Identify and justify the datatype for the below: CO1 BL2
 - A hospital administration tracks the number of patients admitted in the emergency ward every day for an entire month to analyze workload.
 - A weather monitoring station records hourly temperatures in Celsius across different cities to understand climate variations during the winter season.
 - An online survey collects participants' favorite modes of transport by asking them to select from car, bike, bus, train, or walking.
 - A feedback system asks customers to rate their overall satisfaction with a hotel service as poor, fair, good, very good, or excellent.
- Illustrate the necessary data transformation techniques by loading the following *Employee.csv* dataset using Python. The dataset contains employee details such as Emp_Id, Age, Department, Salary, Experience, and Performance_Score. CO2 BL3

Emp_Id	Age	Department	Salary	Experience	Performance_Score
101	25	IT	32000	2	45
112	29	HR	NA	4	60
134	31	Finance	41000	NA	72
146	28	HR	28000	3	35
115	35	IT	50000	7	88
126	26	Finance	33000	2	54
174	30	HR	NA	5	61
138	27	IT	36000	3	73
129	33	Finance	NA	6	85
146	28	HR	28000	3	35
120	NULL	HR	40000	5	NaN
111	29	Finance	45000	4	77
125	24	HR	27000	1	39
134	32	IT	NA	7	90
124	28	HR	31000	3	42
105	36	Finance	55000	10	95

Write a python code to display the output for the following:

- Load the dataset and handle both de-duplication and missing values.
- Convert categorical columns into numerical codes.
- Normalize the Salary and Experience columns.
- Create the following buckets for Performance_Score as follows: Low (0–40), Medium (41–70), High (71–100).

3. Transform the given dataset (V) using the following normalization techniques: min-max normalization between $[-3, 3]$, z-score normalization using mean and mean absolute deviation, and decimal scaling normalization.

Dataset (V): 120, 300, 750, 1100, 1500, 1900

Display the transformation result in a single table.

CO2 BL3

An e-commerce company wants to better understand its customer base and product performance to improve marketing strategies. The company has collected customer demographics and product sales data. The dataset contains customer IDs, age, gender, annual income, product categories, purchase amounts, and purchase dates. Apply Exploratory Data Analysis techniques to extract meaningful insights.

CO3 BL3

Cus_ID	Age	Gender	Annual_Income	Product_Category	Purchase_Amount	Purchase_Date
1001	25	Male	50000	Electronics	250	10-01-2024
1002	NaN	Female	60000	Clothing	120	15-01-2024
1003	45	Male	80000	Home & Kitchen	300	20-02-2024
1004	35	Other	55000	Electronics	450	NaN
1005	22	Female	48000	Clothing	100	22-03-2024
1006	30	Male	70000	Sports & Outdoors	220	25-03-2024
1007	40	Female	90000	Electronics	520	02-04-2024
1008	28	Male	65000	Clothing	180	05-04-2024
1009	50	Female	120000	Jewelry	800	12-04-2024
1010	38	Male	75000	Home & Kitchen	350	18-04-2024

Write a Python code to perform:

- Univariate analysis on the variable *Annual_Income*. Generate descriptive statistics, visualize the distribution using a histogram, and interpret whether the income distribution is normal, skewed, or contains outliers.
- Multivariate analysis using *Annual_Income*, *Product_Category*, and *Purchase_Amount*. Conduct correlation and cross-tabulation to study relationships, generate a grouped bar plot, and interpret how income levels influence purchase behavior across different product categories.

- A data architect working for a retail chain is tasked with analyzing sales performance across multiple regions to support inventory management and seasonal planning. The dataset, *retail_sales.csv*, containing three columns: Date (dd-mm-yyyy) representing daily records over the past 10 years, Region (East/ West/ North/ South), and Sales (total revenue in INR).

CO3 BL4

Describe the key characteristics of a time series dataset, build a comprehensive time series analysis and forecasting model. Import the dataset into Python and derive the insights by implementing time-based indexing, creating basic visualizations, grouping the data by region, applying resampling techniques, and comparing sales trends across different regions.

Dataset: *retail_sales.csv*

Date	Region	Sales (in lakhs)
01-01-2015	North	159
02-01-2015	East	1101
...	...	...
30-12-2024	East	565
31-12-2024	South	658

6. A retail store has collected data on the daily sales (in units) of a product over the last 20 days: CO3 BL4
120, 135, 140, 150, 160, 175, 130, 145, 170, 180, 155, 165, 125, 138, 142, 148, 152, 158, 162, 178
Construct a frequency distribution table for the sales using class intervals of size 10 from the unit starting at 120. Calculate the relative frequency and cumulative frequency for each class. Create a bar plot and a histogram to visualize the distribution. Interpret the histogram and describe whether sales show a normal trend, skewed trend, or multiple peaks.
7. Compute the Singular Value Decomposition of matrix M, $M = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$ and explain CO5 BL4
their significance. Using the SVD, compute the rank of matrix M.
8. A company is developing an autonomous warehouse robot to sort packages efficiently. CO5 BL3
Two teams adopt different learning approaches. Team A trains the robot using a large labeled dataset containing package types and their correct sorting zones. Team B, on the other hand, allows the robot to learn through interaction in a simulated warehouse environment, where it receives a +10 reward for correct sorting, -5 penalty for incorrect sorting, and -2 penalty for collisions or idle behavior. After several simulation rounds, Team B observes that the robot's performance improves as its total reward increases.

Based on this scenario, identify which team is using supervised learning and which is using reinforcement learning, and justify. Discuss how each team's robot receives

feedback during the learning process. Explain why reinforcement learning (Team B's method) might be more suitable for robots operating in real-world environments. Finally, interpret what an updated Q-value represents in simple terms regarding the robot's learning progress.

- 9.a) A logistics company wants to design the most cost-efficient delivery network CO4 BL4 warehouses located in different parts of a region. Each warehouse can be directly connected to some others, and the company has estimated the cost (in thousands of INR) of laying transport routes between them.

From	To	Cost
Northport	Eastville	14
Northport	Westfield	10
Northport	Southbay	18
Eastville	Southbay	12
Eastville	Hilltop	15
Westfield	Southbay	8
Westfield	Riverside	11
Southbay	Riverside	7
Hilltop	Riverside	10
Hilltop	Greenwood	6
Riverside	Greenwood	9
Greenwood	Lakeside	5
Riverside	Lakeside	13

Using Kruskal's algorithm, construct the Minimum Spanning Tree for this warehouse network and show the step-by-step inclusion of edges. Find the total minimum cost for connecting all warehouses. The company decides to create two delivery zones, identify the two clusters of warehouses and briefly explain how this division could improve efficiency in delivery scheduling.

OR

- 9.b) A researcher wants to group documents based on cosine similarity. Consider the following three documents (D1, D2, D3) and one query document (Q): CO4 BL4

D1: "Healthcare systems use AI for diagnosis."

D2: "AI supports healthcare diagnosis and treatment planning."

D3: "Deep learning aids healthcare research."

Q: "AI supports healthcare diagnosis."

Construct a document-term matrix and compute the cosine similarity between the query (Q) and each document (D1, D2, D3). Finally, arrange the documents in descending order of their similarity to Q.

- 10.a) A consumer study measured dissimilarity (1 = very similar, 9 = very dissimilar) among four beverage brands based on taste, price, and packaging. The dissimilarity of D is given below CO5 BL4

$$D = \begin{bmatrix} 0 & 3 & 7 & 6 \\ 3 & 0 & 5 & 4 \\ 7 & 5 & 0 & 2 \\ 6 & 4 & 2 & 0 \end{bmatrix}$$

Apply Classical Metric Multi-Dimensional Scaling to project the brands into a 2D space and plot the resulting 2D coordinates and label each brand.

OR

- 10.b) A device team wants to estimate battery life (hours) from daily screen time (hours). CO5 BL4

Screen Time (hrs)	2	3	4	5	6	7
Battery Life (hrs)	19.5	17.0	14.5	12.0	10.0	5.0

Analyze the process of constructing a Simple Linear Regression Model and predict battery life for Screen Time in 4.5 hours.

⇔⇔⇔ X/K/TY ⇔⇔⇔