RollNo.

ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. /B.Tech / B. Arch (Full Time) – END SEMESTER EXAMINATIONS, NOV/DEC 2024

COMMON TO ALL BRANCHES
Seventh Semester
CSM512 EXPLORATORY DATA ANALYSIS
(Regulation 2019)

Time : 3hrs Max. Marks: 100

CO 1	To outline an overview of exploratory data analysis.
CO 2	To implement data visualization using Matplotlib.
CO 3	To perform univariate data exploration and analysis.
CO 4	To apply bivariate data exploration and analysis.
CO 5	To use Data exploration and visualization techniques for multivariate and time series data.

BL – Bloom's Taxonomy Levels
(L1-Remembering, L2-Understanding, L3-Applying, L4-Analysing, L5-Evaluating, L6-Creating)

PART- A (10x2=20Marks)
(Answer all Questions)

Q. No.	Questions	Marks	CO	BL
1	What is Imputation? Explain with an example.	2	1	2
2	What are non-parametric tests? When are they used?	2	1	2
3	List different visualization techniques for univariate analysis.	2	2	1
4	Draw and explain the components of a box plot.	2	2	2
5	Define Cohen's *d* and its application.	2	3	2
6	What are the factors that affect p-value?	2	3	4
7	Why might rejecting a null hypothesis in a test of a correlation coefficient not be that meaningful? Explain.	2	4	4
8	Briefly explain the different ways of modelling interactions in ANOVA.	2	4	3
9	What are the components of a time series? Give Example.	2	5	3
10	What is stationarity of time series. Why is it important?	2	5	2

PART- B (5x 13=65Marks)
(Restrict to a maximum of 2 subdivisions)

Q. No.	Questions	Marks	CO	BL
11 (a) (i)	Explain different types of data and measurement scales with appropriate examples.	9	1	3
(ii)	Write a Python code snippet to create a dataframe from a .csv file.	4	1	3
	OR			
11 (b)	You are given the following dataset (Table 1) containing customer information for an online retail company with 1000 records. Identify potential issues and propose appropriate cleaning methods for each identified issue. Apply your proposed cleaning methods to the dataset using Python and explain your approach for each transformation.	13	1	3

Table 1 Sales dataset - sample

Cust_ID	Name	Age	Gender	Email	Purchase_amt	Date	Discount
1001	Alice Smith	30	F	alice.smith@email.com	NaN	2024-05-20	Y
1002	Bob Brown	-5	M	bob@email.com	500	2024-06-10	N
1003	Charlie Johnson	45	F	cj@email.com	-20	2024-07-08	N

12 (a)	Given a dataset with the following variables: • **Order ID** (Unique Identifier, type: String) • **Region** (type: String, Values: "North," "South," "East," "West") • **Order Value** (type: float) • **Customer Satisfaction Score** (type: Integer, Values: 1–10) • **Payment Mode** (type: String, Values: "Credit Card," "Debit Card," "UPI," "Cash on Delivery") • **Discount Applied** (type: String, Values: "Yes," "No") 1. Choose the most appropriate plot for univariate analysis of each variable in the dataset. 2. Justify your choice of plots. 3. Write a Python code snippet for generating a univariate plot for any one of the variables.	**13**	2	2
	OR			
12 (b)	With suitable examples discuss the different types of plots available in Matplotlib for univariate and bivariate analysis. Give the code snippet for generating the plots.	**13**	2	3
13 (a)	Critically evaluate the effectiveness of various univariate data exploration techniques (e.g., histograms and summary statistics) in understanding the distribution and central tendencies of a dataset. How would you decide which technique to use based on the characteristics of the data. Provide a detailed example with a dataset of your choice and justify your analysis.	**13**	3	5
	OR			
13 (b)	Explain the different types of t-tests with reference to real-world scenarios. Critically assess the conditions under which each t-test is applicable. Also, demonstrate the application of each type of t-test using Python on a dataset of your choice, and interpret the results.	**13**	3	5
14 (a)	Explain the concept of a contingency percentage table and its use in statistical analysis. What are the different types of contingency tables? Provide an example to illustrate your explanation.	**13**	4	4
	OR			

14 (b)	Perform MANOVA Analysis on the data given in Table 2. Calculate Total Sum of Squares and Between group and Within group Cross Products Matrix. Show step-by-step calculation.	**13**	4	4

Table 2 – Effect of exercise regimen on strength and cardio scores

Exercise Regimen	Strength score	Cardio score
A	70	75
A	72	78
A	68	72
B	85	90
B	88	92
B	87	89

15 (a)	A transportation company wants to analyze daily vehicle traffic data using Python. The dataset consists of two columns: Date (String) and Vehicle_Count (Integer). Write Python code to: • Visualize the time series of vehicle counts. • Check whether the time series is stationary. • Remove trend from the series (detrend). • Compute and interpret the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF).	**13**	4	4
	OR			
15 (b)	Explain the ARIMA model in detail. Analyze the components of ARIMA and their roles in modeling time series data. Provide an example of when and why you would apply each component in a real-world time series forecasting scenario. Additionally, discuss how the model's parameters are selected and their impact on the model's performance.	**13**	4	4

PART- C (1x 15=15Marks)

(Q.No.16 is compulsory)

Q. No.	Questions	Marks	CO	BL
16. (i)	Develop a complete workflow for analyzing the effects of three different marketing strategies on customer purchase amounts. 1. Formulate null and alternative hypotheses. 2. Perform a one-way ANOVA test and explain its assumptions. 3. Explain post hoc tests that can be conducted to identify which marketing strategies differ significantly.	**10**	3	6
(ii)	Explain the differences between Pearson and Spearman correlation coefficients.	**5**	3	4

*****