OR	iii.	What are different techniques to test the assumptions involved in EDA? Identify the strength and weakness of them.	**8**	2	4	4	3
Q.4	i.	What are the issues related with data access in EDA?	**3**	1	2	3	2
	ii.	How to handle missing numerical and non-numerical data in EDA?	**7**	3	3	4	3
OR	iii.	Explain moderate correlation, strong correlation, autoregressive correlation, and sinusoidal correlation.	**7**	4	4	4	3
Q.5	i.	Use the frequency table to make a histogram.	**4**	2	2	3	2

Number of Sit-ups Students Can Do in 1 Minute					
Number	0 – 9	10 – 19	20 – 29	30 – 39	40 +
Frequency	10	8	9	6	2

	ii.	Daniel and 7 of his classmates took the ACT test, a college entrance exam. The student's test scores were as follows: 26, 28, 32, 19, 23, 24, and 31. Using the data provided, make a stem-and-leaf plot. Explain the steps you used.	**6**	3	3	4	3
OR	iii.	Explain Weibull Plot and its importance in manufacturing industry production.	**6**	3	4	4	4
Q.6		Attempt any two:					
	i.	Explain scatter plot and its causal interpretations.	**5**	2	2	3	2
	ii.	Write a short note on Chi-Square calculations, and Phi Coefficient.	**5**	2	4	4	3
	iii.	How do you measure association between mixed combination of numerical, ordinal and nominal variables?	**5**	3	4	4	3

******

Total No. of Questions: 6 *Total No. of Printed Pages:4*

Enrollment No.....

Faculty of Engineering
End Sem Examination Dec 2024
OE00075 Exploratory Data Analytics

Programme: B.Tech. Branch/Specialisation: All

Duration: 3 Hrs. **Maximum Marks: 60**

Note: All questions are compulsory. Internal choices, if any, are indicated. Answers of Q.1 (MCQs) should be written in full instead of only a, b, c or d. Assume suitable data if necessary. Notations and symbols have their usual meaning.

			Marks	BL	CO	PO	PSO
Q.1	i.	What are panel data? (a) Data containing units measured at different time points (b) Data where each unit is measured at more than one time point (c) Data containing skewed variable distributions (d) Data measured at one point in time	**1**	1	1	1	1
	ii.	What characterizes time-series cross-section data? (a) When we have a large number of units recorded at few time points (b) When we have a small or medium sized number of units recorded at many time points (c) When we have a large number of units recorded at many time points (d) When we have a small or medium sized number of units recorded at few time points	**1**	1	1	1	1
	iii.	Which of the following gave rise to need of graphs in data analysis? (a) Data visualization (b) Communicating results (c) Decision making (d) All of these	**1**	1	1	1	1

iv.	Which of the following is characteristic of exploratory graph? (a) Made slowly (b) Axes are not cleaned up (c) Color is used for personal information (d) All of these	**1**	1	1	1	1
v.	Which of the following statements is true about the correlational analysis between two sets of data? (a) The correlational analysis between two sets of data is known as a simple correlation (b) The correlational analysis between two sets of data is known as multiple correlation (c) The correlational analysis between two sets of data is known as partial correlation (d) None of these	**1**	1	1	1	1
vi.	Which of the following statements is true for correlation analysis? (a) It is a bivariate analysis (b) It is a multivariate analysis (c) It is a univariate analysis (d) Both (a) and (c)	**1**	1	1	1	1
vii.	A florist records the amount of money he spent on gasoline each week to deliver orders. The list shows the data for 16 weeks. 32, 38, 40, 44, 44, 46, 49, 53, 55, 58, 62, 65, 68, 70, 72, 81 He put the data in a stem-and-leaf plot. What number did he leave out of the stem-and-leaf plot? (a) 81 (b) 82 (c) 80 (d) 84	**1**	3	3	2	2
viii.	IQ scores are normally distributed with mean 100 and standard deviation 15. What is the IQ scores that correspond to the 85 percentile? (a) 115.6 (b) 120.3 (c) 126.6 (d) 146.2	**1**	3	3	2	2

ix.	Which of the following cannot be covered under univariate analysis of data? (a) Association between two variables (b) Computation of mean, median and mode (c) Preparation of frequency table (d) Computation of percentage frequency for a variable	**1**	2	3	2	2
x.	A third variable is introduced in the two variable table to- (a) Refine the association that was observed originally between two variables (b) The introduction of third variable may show that there was no association between the original two variables (c) Introducing a third variable may indicate association between two original variables although initially no relationship was found between them (d) All of these	**1**	2	3	2	2
Q.2 i.	Give an example of an application of statistics in different business scenarios.	**2**	1	1	1	1
ii.	What is role of population, small sample and large sample in exploratory data analysis?	**3**	1	2	3	2
iii.	Perform classification of data based on different parameters. Give an example in each category.	**5**	2	4	4	3
OR iv.	How exploratory data analysis is performed when data is present in structured, semi structured and unstructured form?	**5**	3	4	4	3
Q.3 i.	What are basic EDA assumptions? What is need of it?	**2**	1	2	3	2
ii.	Differentiate between EDA with classical and EDA with Bayesian Analysis. Give suitable examples.	**8**	3	4	4	3

MarkingScheme

OE00075(T) Exploratory data analysis

Q.1	i.	a) Data containing units measured at different time points	**1**
	ii.	c) When we have a large number of units recorded at many time points	**1**
	iii.	a) Data visualization	**1**
	iv.	c) Color is used for personal information	**1**
	v.	a)The correlational analysis between two sets of data is known as a simple correlation	**1**
	vi.	a)It is a bivariate analysis	**1**
	vii.	b. 82 c. 80 d. 84 are correct	**1**
	viii.	c. 126.6	**1**
	ix.	a. Association between two variables	**1**
	x.	d. All of the above are true.	**1**
Q.2	i.	Give an example of an application of statistics in different business scenarios. **2 example minimum =2 mark**	**2**
	ii.	What is role of population 1M small sample 1M large sample in Exploratory data analysis 1M	**3**
	iii.	Perform classification of data based on different parameters. 3M Give an example in each category. 2M	**5**
OR	iv.	How Exploratory Data Analysis is performed when data is present instructured, semi structured and Unstructured form. **Performance-2 marks** **Parameters- 3marks**	**5**
Q.3	i.	What are basic EDA Assumptions? 1M What is need of it. 1M	**2**
	ii.	Differentiate between EDA with classical and EDA with Bayesian Analysis. Give suitable examples. **EDA with Bayesian 4M, EDA with classical 4M**	**8**
OR	iii.	What are different techniques to test the assumptions involved in EDA. 4M Identify the strength and weakness of them. 4M	**8**
Q.4	i.	What are the issues related with data access in EDA. **Any three issues**	**3**
	ii.	How to handle missing numerical **types of missing data 3M** and non-numerical data in EDA. **4M**	**7**
OR	iii.	Explain moderate correlation, 1.5M strong correlation, 1.5M autoregressive correlation 2M sinusoidal correlation 2M **diagram mandatory**	**7**
Q.5	i.	Use the frequency table to make a histogram.	**4**

Number of Sit-ups Students Can Do in 1 Minute					
Number	0 – 9	10 – 19	20 – 29	30 – 39	40 +
Frequency	10	8	9	6	2

Histogram Diagram -3M Define histogram -1M

	ii.	Daniel and 7 of his classmates took the ACT test, a college entrance exam. The student's test scores were as follows: 26, 28, 32, 19, 23, 24, and 31. Using the data provided, make a stem-and-leaf plot. Explain the steps you used. **Diagram and table required**	**6**
OR	iii.	Explain Weibull Plot 2M and its importance in manufacturing industry production 4M	**6**
Q.6		Attempt any two:	
	i.	Explain scatter plot 2M and its causal interpretations 3M	**5**
	ii.	Write a short note on Chi-Square calculations 2.5 M Phi Coefficient 2.5M	**5**
	iii.	How do you measure association between mixed combination of **Numerical-1M ordinal-2M and nominal variables-2M**	**5**

******

P.T.O.