Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End semester Examination, May 2024

Course: Data Preparation **Semester : IV**
Program: BBA ABD **Time : 03 hrs.**
Course Code: DSIT 2014 **Max. Marks: 100**

Instructions: Attempt all sections

SECTION A
10Qx2M=20Marks

S. No.		Marks	CO
Q 1	Attempt all Questions in this section		
a.	Charts that are helpful in making comparisons are: i. Bar charts ii. column charts iii. Pie charts iv. Both Bar & Column Charts	**2**	**CO1**
b.	What is secondary data? i. Data that isn't as good ii. Data that is collected first-hand iii. Data expressed through interpretive analysis. iv. Data that already exists	**2**	**CO1**
c.	What is data visualization? i. It is the graphical representation of information and data ii. It is the numerical representation of information and data iii. It is the character representation of information and data iv. None of the above	**2**	**CO1**
d.	Data analysts should normalize their numeric variables to: i. Reduce to mean of the data. ii. To convert them into categorical values iii. To create flag variables iv. Standardize the scale of effect each variable has on the results.	**2**	**CO1**
e.	Which one of them is not a method for Outlier Detection? i. Sorting Your Data ii. Boxplots, histograms, and scatterplots iii. Using time series analysis iv. Using the Interquartile Range	**2**	**CO1**
f.	The Z-score method for identifying outliers states that i. Data value is an outlier if it has a Z-score that is either less than −3 or greater than 3.	**2**	**CO1**

	ii. Data value is an outlier if it has a Z-score that is either less than −1 or greater than 1. iii. Data value is an outlier if it is more than 3 times of the mean value iv. Data value is an outlier if it is 5 times the lowest data point.		
g.	The statistical data are of two types. These types are : i. technical data and presentation data ii. Primary data and secondary data iii. Primary data and personal data iv. none of the above	**2**	**CO1**
h.	A graph that uses vertical bars to represent data is called as i. Line graph ii. Bar graph iii. Scatterplot iv. Vertical graph	**2**	**CO1**
i.	Using inter quartile range for identifying outlier, a data value is an outlier if: i. It 2 times the difference between Q3 and Q1 ii. it is located 1.5(IQR) or more below Q1 iii. it is 1.5 times the difference between IQR and mean	**2**	**CO1**
j.	____ are used when you want to visually examine the relationship between two quantitative variables. i. Bar graph ii. pie graph iii. line graph iv. Scatterplot	**2**	**CO1**
SECTION B **4Qx5M= 20 Marks**			
	Attempt all four Questions in this section		
Q.2.	What do you understand by data amputation? Explain with examples	**5**	**CO2**
Q.3.	What do you understand by data cleaning and data transformation? Why are they important?	**5**	**CO2**
Q.4.	What do you understand by binning of numerical variables?	**5**	**CO2**
Q.5.	What is misclassification of data? How do you identify it?	**5**	**CO2**
SECTION-C **3Qx10M=30 Marks**			
	Attempt all three Questions in this section		
Q.6.	Use the following stock price data (in dollars) for following questions. 10 7 20 12 75 15 9 18 4 12 8 1 a. Find the min-max normalized stock price for the stock worth $20. b. Find the Z-score standardized stock price for the stock worth $20. c. Calculate the skewness of the stock price data	**10**	**CO3**
Q.7.	Use the above stock price data for the following. a. Identify the outlier.	**10**	**CO3**

	b. Verify that this value is an outlier, using the Z-score method. c. Verify that this value is an outlier, using the IQR method.		
Q.8.	What are the common methods for binning numerical predictors? Which of these are preferred? Use the following data set for the questions below: 1 1 1 3 3 7 a. Bin the data into three bins of equal width (width = 3). b. Bin the data into three bins of two records each. c. Clarify why each of the binning solutions above are not optimal.	**10**	**CO3**
	SECTION-D **2Qx15M= 30 Marks**		
	Attempt both the Questions in this section		
Q.9.	Consider the given dataset named as iris. Write at least 10 steps for how would you analyze this data in R programing language. Sepal.Length, Sepal.Width, Petal.Length, Petal.Width, Species 5.1, 3.5, 1.4, 0.2, setosa 4.9, 3.0, 1.4, 0.2, setosa 4.7, 3.2, 1.3, 0.2, setosa 4.6, 3.1, 1.5, 0.2, setosa 5.0, 3.6, 1.4, 0.2, setosa 5.4, 3.9, 1.7, 0.4, setosa 4.6, 3.4, 1.4, 0.3, setosa 5.0, 3.4, 1.5, 0.2, setosa 4.4, 2.9, 1.4, 0.2, setosa 4.9, 3.1, 1.5, 0.1, setosa 5.4, 3.7, 1.5, 0.2, setosa	**15**	**CO4**
Q.10.	Consider the following dataset and write commands in R programming language for the following queries. a. Create and explain regression model for LungCap (y), Age (x1) and, Height (x2) (4 marks) b. Create a subset of the given dataset that has all the teenage males who do not smoke. (5 marks) c. Find the correlation between the Age and the Lung Capacity with 99% confidence interval (2 marks)	**15**	**CO4**

d. Create a gender wise boxplot for the Lung Capacity (4 marks)

LungCap	Age	Height	Smoke	Gender	Disease
6.475	6	62.1	no	male	no
10.125	18	74.7	yes	female	no
9.55	16	69.7	no	female	yes
11.125	14	71	no	male	no
4.8	5	56.9	no	male	no
6.225	11	58.7	no	female	no
4.95	8	63.3	no	male	yes
7.325	11	70.4	no	male	no
8.875	15	70.5	no	male	no
6.8	11	59.2	no	male	no
11.5	19	76.4	no	male	yes
10.925	17	71.7	no	male	no
6.525	12	57.5	no	male	no
6	10	61.1	no	female	no
7.825	10	61.2	no	male	no
9.525	13	63.5	no	male	yes
7.875	15	59.2	no	male	no
5.05	8	56.1	no	male	no
7.025	11	61.2	yes	female	no
9.525	14	70.6	no	female	no