Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, Dec 2021

Course: Statistics for Data Science **Semester: V**
Program: B.Tech CSE-SPZ-BD **Time: 03 hrs.**
Course Code: CSBD3006P **Max. Marks: 100**

Instructions:

SECTION A

1. Each Question will carry 4 Marks
2. Instruction: Write short answers for the following questions. (60-70 words)

S. No.		Marks	CO
Q1	Explain confounding variables and elaborate their role in correlation?	**4**	**CO4**
Q2	Mean, median and mode are three primary measures of central tendency. Explain and illustrate the effect of outliers on mean and median?	**4**	**CO1**
Q3	Explain the relation between logarithmic and exponential functions?	**4**	**CO3**
Q4	Discuss the major data types (levels of measurements) used in statistics. Explain briefly the preciousness of these data types?	**2** **2**	**CO2**
Q5	Discuss the differences between dependent and independent variables. Also explain the meaning of lurking variables?	**3** **1**	**CO5**

SECTION B

1. Each question will carry 10 marks.
2. Instruction: Write short / brief notes (100-150 words)
3. For question 6 choose between part a and b.
4. Attempt any one question for question 6
5. There is no such option for other questions in this section

Q6	a. Explain the basics steps in a research process in detail with a suitable example. **OR**	**10**	**CO3**
	b. Serum Institute of India specializing in vaccine states that its Covishield vaccine failure rate is not more than 1%. You perform a hypothesis test to determine whether		

Please Turn Over

	the company's claim is false. When will a type I or type II error occur? Which is more serious? • State the null and alternative hypotheses. • Write the possible type I and type II errors. • Determine which error is more serious.	**3** **3** **4**	
Q7	A group of UPES students were given a short course in speed-reading. The instructor was curious if a monetary incentive would influence performance on a reading test taken at the end of the course. Half of the students were offered Rs 500 for obtaining a certain level of performance on the test, the other half were not offered money. Identify the independent variable and dependent variable in the study.	**10**	**CO1**
Q8	Calculate the correlation coefficient for the gross domestic products and Carbon dioxide emissions data given in the table below. **GDP (Trillions of $), x** / **CO2 Emission (Millions of Metric tons), y** 1.6 / 428.2 3.6 / 828.8 4.9 / 1214.2 1.1 / 444.6 0.9 / 264.0 2.9 / 415.3 2.7 / 571.8 2.3 / 454.9 1.6 / 358.7 1.5 / 573.5 Also display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation.	**5** **5**	**CO4**
Q9	Elaborate the role of clustering in data analytics? Explain the types of data used in cluster analysis.	**3** **7**	**CO5**

SECTION-C

1. Each Question carries 20 Marks.
2. Instruction: Write long answer. (Up to 350 words while explaining)
3. For question 10 choose between part a and b
4. Attempt any one question for question 10.
5. There is no such option for other question 11.

Q10.	**a.** Find the mean, the median, and the mode of the sample ages of students in a class shown at the left. Which measure of central tendency best describes a typical entry of this data set? Are there any outliers? Age of students in a class: 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 22, 22, 22, 23, 23, 23, 23, 24, 24, 65 • With the help of histogram display the distributions of data along with locations of mean, median, mode. • Remove the data entry 65 from the data set and then calculate the mean, median and the mode. Does the absence of the outlier change the measures? If yes, justify your answer.	**8** **6** **6**	**CO2**
	OR **b.** The following sample data set lists the per hour salaries of 20 employees. Construct the five-point summary of the sample data set The sample data is as follows: 30, 32, 110, 65, 55, 50, 55, 48, 43, 42, 45, 45, 33, 34, 38, 38, 34, 38, 33, 32 Analyze the data set and apply the five-point summary methods to calculate the following: • First quartile, Second quartile, Third quartile • Range • Interquartile range • Semi interquartile range • Detect the outliers if any • Construct a box plot for displaying the five-point summary	**3** **3** **3** **3** **3** **5**	
Q11.	The regressions line always passes through the averages of data points. The formula for calculating the slop of the regression line is m = r (sy/sx) where m is the slope, r is the regression coefficient, sy and sx are standard deviations.		**CO4/ CO5**
	a. Using this information prove that the slope m = ∑(yi – y-mean)/ ∑(xi – x-mean) where yi and xi are ith observation for x and y respectively.	**12**	
	b. Use the data set in Question 8 to experimentally evaluate that r (sy/sx) = ∑(yi – y-mean)/ ∑(xi – x-mean).	**8**	