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

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
Online End Semester Examination, Dec 2020

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 5 Marks
2. Instruction: Complete the statement / Select the correct answer(s)

S. No.		Marks	CO
Q1	The set of all possible sample points (experimental outcomes) is called Select one: a) a sample b) an event c) sample space d) All of the above	**5**	**CO1**
Q2	The entities on which data are collected are a) variables b) datasets c) elements d) none of the above	**5**	**CO2**
Q3	If the assumed hypothesis is tested for rejection considering it to be true is called? a) Null Hypothesis b) Statistical Hypothesis c) Simple Hypothesis d) Composite Hypothesis	**5**	**CO3**
Q4	The most frequently occurring data value in a data set is the a) Median b) Arithmetic mean c) Range d) Mode	**5**	**CO4**

Q5	The probability distribution for a discrete random variable that is used to compute the probability of x occurrences of an event over a specified interval is known as a) the normal probability distribution b) the standard normal distribution c) a discrete random variable d) poisson probability distribution	**5**	**CO5**
Q6	Statistics branches include a) Applied Statistics b) Mathematical Statistics c) Industry Statistics d) Both a and b	**5**	**CO1**
SECTION B **1. Each question will carry 10 marks** **2. Instruction: Write short / brief notes**			
Q7 (a)	Explain the basics steps in a research process in detail with a suitable example.	**5**	**CO1**
(b)	Identify the independent variable and dependent variable in the following studies. 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 the students were offered Rs 500 for obtaining a certain level of performance on the test, the other half were not offered money.	**5**	
Q8	The following sample data set lists the prices (in dollars) of 30 portable global positioning system (GPS) navigators. Construct a frequency distribution that has seven classes. The sample data is as follows: 90 130 400 200 350 70 325 250 150 250 275 270 150 130 59 200 160 450 300 130 220 100 200 400 200 250 95 180 170 150 Using the frequency distribution constructed, calculate the • Midpoints. • Relative Frequency. • Cumulative Frequency.	**10**	**CO2**
Q9	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 the company's claim is false. When will a type I or type II error occur? Which is more serious? a) State the null and alternative hypotheses. b) Write the possible type I and type II errors. c) Determine which error is more serious.	**10**	**CO3**

<table>
<tr><td>Q10</td><td>Calculate the correlation coefficient for the gross domestic products and Carbon dioxide emissions data given in the table below. Also display the data in a scatter plot and determine whether there appears to be a positive or negative linear correlation.

<table><tr><th>GDP(Trillions of $),x</th><th>CO2 Emission(Millions of Metric tons),y</th></tr><tr><td>1.6</td><td>428.2</td></tr><tr><td>3.6</td><td>828.8</td></tr><tr><td>4.9</td><td>1214.2</td></tr><tr><td>1.1</td><td>444.6</td></tr><tr><td>0.9</td><td>264.0</td></tr><tr><td>2.9</td><td>415.3</td></tr><tr><td>2.7</td><td>571.8</td></tr><tr><td>2.3</td><td>454.9</td></tr><tr><td>1.6</td><td>358.7</td></tr><tr><td>1.5</td><td>573.5</td></tr></table></td><td>10</td><td>CO4</td></tr>
<tr><td>Q11</td><td>What is clustering? Explain the types of data used in cluster analysis.</td><td>10</td><td>CO5</td></tr>
<tr><td colspan="4">SECTION-C
1. Each Question carries 20 Marks.
2. Instruction: Write long answer.</td></tr>
<tr><td>Q12.</td><td>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?

<table><tr><td>Age of students in a class</td><td>20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 22, 22, 22, 23, 23, 23, 23, 24, 24, 65</td></tr></table>
With the help of histogram display the distributions of data along with locations of mean, median, mode.

b) 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.</td><td>20</td><td>CO2</td></tr>
<tr><td></td><td></td><td></td><td></td></tr>
</table>