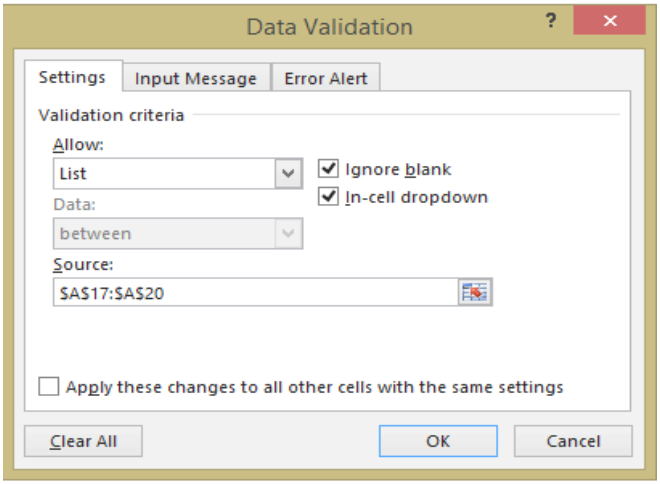
Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

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

Course: Data Preparation — **Semester: V**
Program: BBA Analytics and Big Data — **Time: 03 Hours**
Course code: DSIT3004 — **Max. Marks: 100**

SECTION A **(20 Marks)**

1.	**Each question in section A is a multiple-choice question with four answer choices. Read each question and choose the one best answer.**	**Marks**	**CO**
i)	Which of the following is not a Tableau Prep field data type: a) Number(Whole) b) String c) Boolean d) Float	**2**	**CO1**
ii)	For creating variable size bin we use ……. a) Sets b) Groups c) Calculated fields d) Table Calculations	**2**	**CO1**
iii)	Missing data isn't really a problem if I'm just doing simple statistics, like chi-squares and t-tests. a) True b) False	**2**	**CO1**
iv)	Based on the data validation rule below, what will show up in the drop down list in the cell?  a) The values shown in cells A17 through A20 b) $A$17:$A$20 c) The values A17, A18, A19, A20 d) Don't know	**2**	**CO1**

v)	What is the file extension for tableau prep flow file: a) .tde b) .tfd c) .tfl d) .tdf	**2**	**CO1**
vi)	In standardization, the features will be rescaled with a) Mean 0 and Variance 0 b) Mean 0 and Variance 1 c) Mean 1 and Variance 0 **d)** Mean 1 and Variance 1	**2**	**CO1**
vii)	To remove noise and inconsistent data ____ is needed. a) Data Cleaning b) Data Transformation c) Data Reduction d) Data Integration	**2**	**CO1**
viii)	Multiple data sources may be combined is called as _____ a) Data Reduction b) Data Cleaning c) Data Integration d) Data Transformation	**2**	**CO1**
ix)	Examples of Ordinal scale can be: a) ID Numbers, eye color, zip codes b) Rankings, taste of potato chips, grades, height c) Calendar dates, temperatures in Celsius or Fahrenheit, phone numbers d) D. Temperature in Kelvin, length, time, counts	**2**	**CO1**
x)	Which of the following is true about outliers - a) Data points that deviate a lot from normal observations b) Can reduce the accuracy of the model c) Both A and B d) None	**2**	**CO1**

SECTION B **(20 Marks)**

This section has 4 Questions of 5 marks each.
These questions are short answer type.
All the questions are compulsory.

2.	Write Five useful spreadsheet file format for data storage.	**5**	**CO2**
3.	What are the three different uses of Calculated filed in tableau Prep.	**5**	**CO2**
4.	To validate few cells in excel for data entry what will be formula if condition is that each entry should be unique start with a fixed letter and length of the entry is 5.	**5**	**CO2**
5.	What is the uses of QQ plot in data preparation.	**5**	**CO2**

SECTION-C **(30 Marks)**

This section has 3 Questions of 10 marks each, out of which first 2 Questions are compulsory.
Questions 8 has internal choice to attempt any one.

6.	What is standard score transformation discuss with example.	**10**	**CO3**
7.	What is the use of Jarque-Bera Test. Discuss with example.	**10**	**CO3**
8.	What are different type of missingness. Discuss with example. **OR** What is the difference between case deletion technique and pairwise deletion techniques. Discuss with example.	**10**	**CO3**

SECTION-D **(30 Marks)**

This section has 2 Questions of 15 marks each, out of which Question 9 is compulsory and Question 10 has internal choice to attempt any one.

9.	For the given data set suggest and apply two methods for imputing the missing values and discuss the merit and demerit of the results obtained by the suggested methods.	**15**	**CO4**

Y	X
10	8
14	7
12	
10	6
8	
9	7

10. Using Grubbs' Test verify that in the following data set there is actual outlier present or not.(Assume that data is normally distributed and use t-critical=3.3)

145,125,190,135,220,130,210,3,165,165,150

OR

Using the Chi-square test verify that the following data is normally distributed or not at 5% Level of significance.

Returns	Observed Frequency
-15% to -5%	22
-5% to +5%	29
+5% to +15%	37
+15% to +25%	12

(Given $P(Z \leq 0) = 0.5, P(Z \leq 1) = 0.1587, P(Z \leq 2) = 0.9773$

15 **CO4**

Chi-square Distribution Table

d.f.	.995	.99	.975	.95	.9	.1	.05	.025	.01
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21