SOMAIYA VIDYAVIHAR UNIVERSITY

Semester: January 2022 – May 2022 Examination: In-Semester Examination		
Programme code: 54 **Programme: B.Tech. Computer Engineering** **(Honours in Data Science and analytics)**	**Class:** SY	**Semester:** IV **(SVU 2020)**
Name of the Constituent College: K. J. Somaiya College of Engineering	**Name of the department:** COMP ENGG	
Course Code: 116h54C401	**Name of the Course: Applied Data Science**	

Question No.		Max. Marks	CO Mapped
Q1	Attempt any FIVE(5) of following A. What are the characteristic of big data? B. List out skillset required for data science profile. C. What is need of data scrapping? D. List types of the exploratory data analysis E. Give an example of binomial distribution F. What is data wrangling?	2 mark Each	CO1 and CO3
Q2	Attempt any TWO (2) of the following a) Illustrate with suitable example how data science can be used to add value to business. b) A company packages salted peanuts in 200gm packets using the machine. A sample of 16 packets is taken from the production line at random time intervals and their contents are weighted. The mean weight of 16 packets is found as 199.5. Can we conclude that machine is working properly using statistical hypothesis testing? In the hypothesis testing clearly show all the 5 steps in the calculation, considering the significance level of 5%. (Assume value of test statistic=2.131) c) Are the following nominal, ordinal, interval or ratio data? Explain your answers. (a) Temperatures measured on the Kelvin scale. (b) Military ranks. (c) Social security numbers. (d) Number of passengers on buses from Delhi to Mumbai. (e) Code numbers given to the religion of persons married.	05 marks 05 marks 05 marks	CO1 CO3 CO1
Q3	a) You have been given the task to perform the data preprocessing of the data retrieved from multiple sources, before you start applying the data mining task. Identify (atleast 5) data quality issues with the sample dataset retrieved from the master data set. Suggest how do you resolve the quality issue.	05 Marks	CO3

TXN-ID	NAME	AGE	HEIGHT	WEIGHT	BLOOD GROUP	COVID-19 RESULT
T001	RAMA	45	145	62kg	O+ve	Positive
T002	SEETHA	43	168	45kg	B+ve	Negative
T003	Akbar	38	172	60kg	Iam+ve	Positive
T004	BIRBAL	45	168	52kg	AB+ve	Negative
T005	THenali	22	157	78kg	B-ve	1
T006	Venkat	36	157	54kg	O-ve	Negative
T007	Rajuu	350	132	48kg	O+ve	Positive
T008	HARI	32	180	120lbs	AB-ve	Negative
T009	Inba	25		85kg	O+ve	0
T010	SysUsr789	20	165	68kg	O-ve	Negative

OR

Eleven students were asked to measure their pulses for 30 seconds and multiply by two to get their one minute pulse rate. The results were: 62,32,60,66,70,72,74,74,78,80,84. Give the five-number summary for the pulse rates and draw the boxplot — 05 Marks — CO1

b) Use min-max normalization method to normalize the following group of data by setting min=0 and max=1:
200, 300, 400, 600,1000 — 03 Marks — CO3

c) What is type I and Type II error? Provide example of each. — 02 Marks — CO3