Roll No												

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF INFORMATION SCIENCES

MAKE UP EXAMINATION – JAN 2023

Date: 23-01-2023

Time: 9:30 am to 12.30 pm

Max Marks: 100

Weightage:50%

Course Code: BSD2002

Course Name: Data Modelling & Visualization

Program : B.Sc(DS)

Instructions:

(i) *Read all the questions carefully and answer accordingly.*

Part A [Memory Recall Questions]

Answer all the Questions. Each question carries TWO marks. (10Qx 2M= 20M)

1. What is geospatial Data? (**C.O.No.4**) [**Knowledge**]

2. Define data visualization. (**C.O.No.4**) [**Knowledge**]

3.List any 2 python libraries required for data visualization. (**C.O.No.4**) [**Knowledge**]

4 What is a pie chart? (**C.O.No.4**) [**Knowledge**]

5. Define Data Science. (**C.O.No.4**) [**Application**]

6. Define unstructured data? (**C.O.No.1**) [**Knowledge**]

7. List any two industries in which data science is useful. (**C.O.No.3**) [**Knowledge**]

8. List any two benefits associated with data extraction. (**C.O.No.2**) [**Knowledge**]

9. List any two reinforcement learning algorithms. (**C.O.No.1**) [**Knowledge**]

10. List any two methods of imputing missing values. (**C.O.No.2**) [**Application**]

Part B [Thought Provoking Questions]

Answer all the Questions. Each question carries TEN marks. (5Qx10M=50M)

11. Explain in detail the common mistakes that data scientists make which leads to ineffective data visualization resulting in wrong business decisions. (**C.O.No.4**) [**Application**]

12. You have been provided with a tourist dataset to Europe from the year 1960 to 2018 which contains year on year number of tourists. Using EDA and necessary python libraries, write at least 15 different steps involved to visualize this type of data.

(**C.O.No.4**) [**Application**]

13. Stroke is one of the main reasons for disability and death. The lifetime risk of an adult person is about 25% of having once a stroke. But stroke is a very heterogeneous disorder. So, having individualized pre-stroke and post-stroke care is crucial for the success of a cure.

To determine this individualized care, the person's phenotype, i.e., observable characteristics of a person should be chosen. And this is usually achieved by biomarkers. A so-called biomarker is a measurable data point such that the patients can be stratified. Examples are disease severity scores, lifestyle characteristics, or genomic properties.

There are many known biomarkers already published or in databases. Also, there are daily many hundreds of scientific publications about the detection of biomarkers for all the different diseases.

Research is enormously expensive and time-critical to prevent a disorder. So, biotech companies need to understand the most effective and efficient corresponding biomarkers for a particular disease. The amount of information is so gigantic that this cannot be done manually.

As a data analyst, please discuss the way you are going to design a model for the above scenario. (**C.O.No.2**) [**Application**]

14. Tesla is now a big player in the electric automobile industry. It is widely known for its advanced and futuristic cars. The company says that the cars have their own AI hardware. Tesla is using AI for making self-driving cars. At the moment, cars are not completely autonomous. The company is working on the thinking algorithm for cars. It is currently working with NVIDIA on an unsupervised ML algorithm.

This step by Tesla would be a game-changer for many reasons. The cars send data directly to tesla's cloud. How data collected in Tesla's cloud can be used in improving the sales of electric automobile industry. (**C.O.No.3**) [**Application**]

15. The production of drugs needs time, especially today's high-tech cures based on specific substances and production methods. Also, the whole processes are broken down into many different steps, and several of them are outsourced to specialist providers.

We see this currently with the COVID-19 vaccine production. The vaccine inventors deliver the blueprint, and the production is in plants of companies specialized in sterile production. The vaccine is delivered in tanks to companies that do the filling in small doses under clinical conditions, and finally, another company makes the supply.

Further, drugs can be stored only for a limited time and often under special storage conditions, e.g., in a cold storage room.

Design a data model in order to process this requirement in an efficient manner.

(**C.O.No.3**) [**Application**]

Part C [Problem Solving Questions]

Answer all the Questions. Each question carries FIFTEEN marks. (2Qx15M=30M)

16.You are provided with XYZ dataset. The first five entries are shown in the figure. List atleast ten commands that you will use to have a better understanding about your data. Suggest a suitable data visualization technique.

	Month	#Passengers
0	1949-01	112
1	1949-02	118
2	1949-03	132
3	1949-04	129
4	1949-05	121

(**C.O.No.4**) [**Application**]

17. Explain any 5 encoding techniques with suitable examples. (**C.O.No.2**) [**Comprehension**]