Roll No												

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF INFORMATION SCIENCE

MID TERM EXAMINATION
(SET A)

Winter Semester: 2021 - 22
Course Code: BSD 2002
Course Name: Data Modeling and Visualization
Program & Sem: BSc(DS) – IInd Sem

Date: 14/05/2022
Time: 10:00 AM – 11:30 AM
Max Marks: 50
Weightage: 25%

Instructions:

(i) Read the all questions carefully and answer accordingly.
*(ii) Please access the required dataset from respective **MS Teams** account.*

Part A [Memory Recall Questions]

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

1. Data science is the process of diverse set of data through? (**C.O.No.1**) [**Knowledge**]
a) Organizing data b) Processing data c) analyzing data d) all of the above

2. Identify the right skills for a Data Scientist? (**C.O.No.1**) [**Knowledge]**
a) Probability & Statistics b) Machine Learning c) Data Wrangling d) all of the above

3. Identify the forms of Data sources in data science? (**C.O.No.1**) [**Knowledge**]
a) Structured b) Unstructured c) Both A and B d) None of the above

4. Identify which of the following is not an application for data science?
(**C.O.No.1**) [**Comprehension**]
a) Recommendation Systems b) Image and Speech Recognition
c) online price comparison d) Privacy checker

5. State true or false. Unstructured data is not organized. (**C.O.No.1**) [**Comprehension**]
a) True b) False c) Can be true or false d) Can't say

6. A column is a _____ representation of data. (**C.O.No.1**) [**Knowledge**]
a) horizontal b) diagonal c) vertical d) top

7. A _____ is a structured representation of data. (**C.O.No.1**) [**Knowledge**]
a) database table b) functions c) data preparation d) data frame

8. Raw data should be processed only one time. (**C.O.No.2**) [**Comprehension**]
a) True b) False c) can be true or false d) can't say

9. Which of the following step is performed by data scientist after acquiring the data?
(**C.O.No.2**) [**Comprehension**]

a) Data Cleaning b) Data Integration c) Data replication d) All of the above

10. Which of the following focuses on the discovery of (previously) unknown properties of the data? (**C.O.No.2**) [**Comprehension**]

a) Data mining b) BIgData c) Data wrangling d) machine learning

Part B [Thought Provoking Questions]

Answer all the Questions. **(2Q x 1M = 02Marks)**
(2Q x 3M = 06Marks)
(6Q x 2M = 12Marks)

11. Define the dataframe, inp0 by reading the dataset of “bank telemarketing campaign”.
(**1 Mark)** (**C.O.No.2**) [**Knowledge**]
12. Identify the last 10 rows of inp0. (**1 Mark)** (**C.O.No.2**) [**Comprehension**]
13. Identify the data types of each column in inp0. (**2 Marks)** (**C.O.No.2**) [**Comprehension**]
14. Modify the inp0 by dropping the customer id of inp0. (**2 Marks)** (**C.O.No.2**) [**Application**]
15. Convert the age variable data type to the appropriate one in inp0.
(**2 Marks)** (**C.O.No.2**) [**Application**]
16. Modify inp0 into inp1 after dropping the records with missing values in age.
(**2 Marks)** (**C.O.No.2**) [**Application**]
17. Calculate the percentage of each marital status category in inp1.
(**3 Marks)** (**C.O.No.2**) [**Application**]
18. Demonstrate the bar graph for percentage marital status categories.
(**3 Marks)** (**C.O.No.2**) [**Application**]
19. Modify the inp1 by dropping the records which have missing values in response.
(**2 Marks)** (**C.O.No.2**) [**Application**]
20. Extract job in newly created ‘job’ column from “jobedu” column.
(**2 Marks)** (**C.O.No.2**) [**Application**]

Part C [Problem Solving Questions]

Answer all the Questions. The question carries 10 Marks **(1Q x 10 = 10Marks)**

21. Explain outliers. Use the Inter Quartile range to find the outliers in the following data.

25, 37, 24, 28, 35, 22, 31, 53, 41, 64

(**C.O.No.1, 2**) [**Application level**]

Roll No												

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF INFORMATION SCIENCES

END TERM EXAMINATION

Winter Semester: 2021 - 22
Course Code: BSD 2002
Course Name: Data Modelling & Visualization
Program & Sem: B.Sc (DS)

Date: 6th July 2022
Time: 01.00 PM to 04.00 PM
Max Marks: 100
Weightage:50%

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

Part A [Memory Recall Questions]

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

1. Define is Time Series Data? (**C.O.No.4**) [**Knowledge**]

2. Define Line Data and Area Data. (**C.O.No.4**) [**Knowledge**]

3.List any 2 visualization techniques for geospatial data. (**C.O.No.4**) [**Knowledge**]

4 List any 2 common mistakes that make data visualization ineffective. (**C.O.No.4**) [**Knowledge**]

5. Explain univariate and multivariate data. (**C.O.No.2**) [**Comprehension**]

6. Define Data Science. (**C.O.No.1**) [**Knowledge**]

7. List any 2 skills that are required for a Data scientist. (**C.O.No.2**) [**Knowledge**]

8. Identify the command in Python to display the last 15 rows of dataframe(df). (**C.O.No.1**) [**Comprehension**]

9. Explain the output of describe() command on dataframe(df). (**C.O.No.2**) [**Comprehension**]

10. Define Data extraction. (**C.O.No.3**) [**Knowledge**]

Part B [Thought Provoking Questions]

Answer all the Questions. Each question carries 10 marks. (5Qx10M=50M)

11. Explain in detail the basic components of a Time Series Data. Also discuss decomposition models of time series data. Discuss the various data wrangling steps required before time series data visualization? (**C.O.No.4**) [**Comprehension**]

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. ABC bank aimed at encouraging its customers to subscribe to terms deposits by calling them and pitching the service. The bank hired you as an analyst and asked to illustrate a model which will help the bank in deciding the called customer will invest in term deposit or not.

(**C.O.No.2**) [**Application**]

14. We have a vast number of micro-organisms, so-called microbiota like bacteria, fungi, viruses, and other single-celled organisms in our body. All the genes of the microbiota are known as the microbiome. The number of these genes is trillions, and, e.g., the bacteria in the human body have more than 100 times more unique genes than humans.

The microbiota has a massive influence on human health, and imbalances are causing many disorders like Parkinson's disease or inflammatory bowel disease. There is also the presumption that such imbalances cause several autoimmune diseases. So, microbiome research is a very trendy research area.

To influence the microbiota and develop microbiome therapeutics to reverse diseases, one needs to understand the microbiota's genes and influence on our body. With all the gene sequencing possibilities today, terabytes of data are available but not yet probed. Illustrate a model to develop microbiome-targeted treatments and predict microbiome-drug interactions.(**C.O.No.3**) [**Application**]

15. Heart failure typically leads to emergency or hospital admission. And with an aging population, the percentage of heart failure in the population is expected to increase.

People that suffer heart failure usually have pre-existing illnesses. So, it is not uncommon that telemedicine systems are used to monitor and consult a patient, and mobile health data like blood pressure, body weight, or heart rate are collected and transmitted.

Most prediction and prevention systems are based on fixed rules, e.g., when specific measurements are beyond a pre-defined threshold, the patient is alerted. It is self-explanatory that such a predictive system has a high number of false alerts, i.e., false positives.

Because an alert leads mostly to hospital admission, too many false alerts lead to increased health costs and deteriorate the patient's confidence in the prediction. Eventually, she or he will stop following the recommendation for medical help.

i) Identify the features that you want to collect from the patients.
ii) Explain how EDA can be used to reduce these false positives.

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

Part C [Problem Solving Questions]

Answer both the Questions. Each question carries 15 marks. (2Qx15M=30M)

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

	area	bedrooms	age	price
0	2600	3.0	20	550000
1	3000	4.0	15	565000
2	3200	NaN	18	610000
3	3600	3.0	30	595000
4	4000	5.0	8	760000
5	4100	6.0	8	810000

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

17. Discuss outliers along with ways of handling them. Using IQR find the outliers in the given dataset.

25	37	24	28	35	22	31	53	41	64	29

(**C.O.No.2**) [**Comprehension**]