NOIDA INSTITUTE OF ENGINEERING AND TECHNOLOGY, GREATER NOIDA

(An Autonomous Institute Affiliated to AKTU, Lucknow)

B.Tech.

SEM: III - THEORY EXAMINATION (2021 - 2022)

Subject: Foundations of Data Science

Time: 03:00 Hours   Max. Marks: 100

General Instructions:

1. All questions are compulsory. It comprises of three Sections A, B and C.

- Section A - Question No- 1 is objective type question carrying 1 mark each & Question No- 2 is very short type questions carrying 2 marks each.
- Section B - Question No- 3 is Long answer type - I questions carrying 6 marks each.
- Section C - Question No- 4 to 8 are Long answer type - II questions carrying 10 marks each.
- No sheet should be left blank. Any written material after a Blank sheet will not be evaluated/checked.

SECTION A

20

1. Attempt all parts:-

1-a.   Which of the following uses data on some object to predict values for other object   1
(CO1)

               1. Inferential

               2. Exploratory

               3. Predictive

               4. None of the mentioned

1-b.   Which of the following step is performed by data scientist after acquiring the data   1
(CO1)

               1. Data Cleansing

               2. Data Integration

               3. Data Replication

               4. All of the mentioned

1-c.   What does the following block of code do?   1

```
url = "https://www.nytimes.com"
html = urllib.request.urlopen(url, context=ctx).read()
soup = BeautifulSoup(html, 'html.parser') (CO2)
```

               1. retrieves and displays the webpage

               2. parses the html content of the "https://www.nytimes.com" webpage.

               3. downloads the webpage

               4. It throws an error because a socket cannot use HTTP

1-d. What is an essential process in which the intelligent methods are applied to extract data? (CO2)   1

      1. Warehousing

      2. Data Mining

      3. Text Mining

      4. Data Selection

1-e. R objects can have attributes, which are like _____ for the object. (CO3)   1

      1. metadata

      2. features

      3. expression

      4. dimensions

1-f. A _____ is a two-dimensional rectangular data set. (CO3)   1

      1. Matrix

      2. Lists

      3. Vector

      4. Functions

1-g. Which of the following is an example of raw data? (CO4)   1

      1. original swath files generated from a sonar system

      2. initial time-series file of temperature values

      3. a real-time GPS-encoded navigation file

      4. all of the mentioned

1-h. Which of the following return a subset of the columns of a data frame? (CO4)   1

      1. select

      2. retrieve

      3. get

      4. set

1-i. Which function is used to create 3D Plot in R? (CO5)   1

      1. range()

      2. matrix()

      3. persp()

      4. pnorm()

1-j. Plot used to show the relationship between two sets of data (CO5)   1

      1. Time line

      2. Scatter Plot

      3. Bubble Chart

      4. None of these

2. Attempt all parts:-

2-a. Explain the process of datafication (CO1)   2

2-b. Describe unstructured data with example (CO2)   2

2-c.	What is the process of loading a .csv file in R? (CO3)	2
2-d.	List main functions of Janitor package (CO4)	2
2-e.	Describe the working of a web scraper (CO5)	2

SECTION B

30

3. Answer any <u>five</u> of the following:-

3-a.	Discuss all phases of Data Science lifecycle (CO1)	6
3-b.	Explain all the components of Hadoop ecosystem (CO1)	6
3-c.	Differentiate between qualitative and quantitative data with examples. Mention their types (CO2)	6
3-d.	What is an outlier? How you detect outliers in your data? (CO2)	6
3-e.	Name some functions available in "dplyr" package. Describe them with examples (CO3)	6
3-f.	Distinguish between dimensionality reduction and numerosity reduction (CO4)	6
3-g.	List down the advantages of data visualization in R (CO5)	6

SECTION C

50

4. Answer any <u>one</u> of the following:-

4-a.	Briefly explain crowd sourcing analytics with example. Also mention its types and cause of its rise in 21st century. (CO1)	10
4-b.	Explain how Uber and Facebook are using data science techniques for data analytics (CO1)	10

5. Answer any <u>one</u> of the following:-

5-a.	(a) What is data normalization? What are the methods of normalizing data? (b) Explain the process of binning with example (CO2)	10
5-b.	What is data preprocessing? Explain the major steps involved in the process with example. (CO2)	10

6. Answer any <u>one</u> of the following:-

6-a.	df<-data.frame(Name=c(NA, 'John', 'Arun', NA, 'Andrew'), 　　　　　　Sales=c(20,18,22,55,59), 　　　　　　Price=c(33,51,20,40,20), 　　　　　　stringsAsFactors=FALSE) Write a R code that will remove all NA from Name Column Write a R code that will remove all NA from entire data frame (CO3)	10
6-b.	How is a factor different from a dataframe? Write a R program to get All Factor Levels of DataFrame Column (CO3)	10

7. Answer any <u>one</u> of the following:-

7-a.	Explain the process of Principal Component Analysis and illustrate with example. How is it different from Linear Discriminant Analysis? (CO4)	10
7-b.	Explain ways to perform Bivariate analysis for Numerical-numerical, Categorical-Categorical, and Numerical-Categorical variables (CO4)	10

8. Answer any <u>one</u> of the following:-

8-a.   How can we visualize spatial data and maps in R? what are the packages available   10
        for spatial data? (CO5)

8-b.   What are the ways of data visualization? Explain how does visualization of big data   10
        help in interpreting information? (CO5)