VIT-AP
UNIVERSITY
Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2023-2024–CAT-1

Course Code: CSE1006 **Course Title: Foundations for Data Analytics**

Set number: 3 **Date of Exam:** 10/02/2024 (FN) (F1)

Duration: 90 Minutes **Total Marks: 50**

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. Assume that you are a data science student and working on a project. Your project is based on weather forecasting analysis. In this regard, data gathered from online sources consisting of different parameters/attributes is shown below:

Date	Max_Temp	Min_Temp	Humidity	Rainfall	Outlook
04-Nov	41	29	75	0	Cloudy
05-Nov	50	35	85	0	Cloudy
06-Nov	46	25	75	3	Cloudy
07-Nov	50	35	79	3	Cloudy
08-Nov	55	40	78	0	Sunny
09-Nov	41	27	75	0	Sunny
10-Nov	55	37	73	0	Sunny
11-Nov	41	27	75	0	Sunny
12-Nov	43	37	62	3	Cloudy
13-Nov	55	40	65	3	Cloudy
14-Nov	58	35	67	0	Cloudy
15-Nov	55	37	84	3	Cloudy
16-Nov	43	37	92	0	Sunny
17-Nov	50	35	90	0	Cloudy

The analysis tasks are:

a) Find the mean of the Max_Temp, Min_Temp, Humidity? **(3M)**

b) Find the median of the Max_Temp, Min_Temp, Humidity? **(3M)**

c) Find the mode of the Min_Temp? **(3M)**

d) Find the Variance of the Max_Temp, Min_Temp, Humidity? **(3M)**

e) Find the standard deviation of the Max_Temp, Min_Temp, Humidity? **(3M)**

Q2. a) Define Data and Information. Distinguish between Data and Information.**(5M)**

b) What is Data scrapping? Describe different types of data scraping.**(5M)**

Q3. Assume that there are 5 people following food diet for weight loss. The weights(in kg) of 5 people before & after a diet is given in the following table

	P1	P2	P3	P4	P5
Before	87	83	81	90	99
After	74	66	65	72	85
Weight_lost					
Gender					

a) Create two vectors named as "Before" and "After" with corresponding values and display them.**(3M)**

b) Evaluate the amount of weight lost by each participant, store it into a vector "Weight_lost", and display the results.**(2M)**

c) Write a function to find each participant's gender whether it is either Male(M) or Female(F). Assume that if weight_lost is>15, then those participants' gender is Female.**(5M)**

Q4. a) Write a R program to create a matrix using the following data:**(5M)**

i. Range of numbers is from 1 to 9
ii. Number of rows is 2 and enables the row wise operation.
iii. Find the dimension of the matrix.

b) Write a R program to find the factorial of an user defined numerical input.**(5M)**

c) Consider the following vector: x <- c(10,12,NA,14,15,16,NA,18,NA,19) and answer the following: **(5M)**

i. Write a command to find the size of the vector and find if it is a vector or not.
ii. Count number of 'NA' values.
iii. Replace all the missing values with the mean of the remaining values and display the updated vector.

QP MAPPING

Q. No.	**Module Number**	**COMapped**	**PO Mapped**	**PEO Mapped**	**PSO Mapped**	**Marks**
Q1	1	1	1, 2, 3	1, 2, 3, 4	1, 3	15
Q2	1	1	1, 2, 3	1, 2, 3, 4	1, 3	10
Q3	2	2	1, 2, 3	1, 2, 3, 4	2	10
Q4	2	2	1, 2, 3	1, 2, 3, 4	2	15

VIT-AP
UNIVERSITY
Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2023-2024–CAT-1

Course Code: CSE1006	**Course Title: Foundations for Data Analytics**
Set number: 4	**Date of Exam:** 08/02/2024 (FN) (D1)
Duration: 90 Minutes	**Total Marks: 50**

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. What is data wrangling? Describe its functionalities for data analysis using the following data.**(15M)**

	Name	Age	Gender	Marks
0	Jai	17	M	76
1	Prince	18	M	NA
2	Guru	18	M	90
3	Anu	17	F	85
4	Raju	17	M	NA
5	Neha	18	F	74
6	Riyaj	17	F	80

Functionalities:

a. Data exploration
b. Dealing with missing values
c. Replacing data
d. Filtering data

Q2. (a) Define the concepts of data and information. Explain the key characteristics of data.**(5M)**

(b) Define the concept of sampling and list-out the different methods. Describe the challenges of data sampling.**(5M)**

Q3. (a) Write a program in R to check whether a number is prime or not. **(10M)**

Sample Output:

```
Input a number to check prime or not: 13
The entered number is a prime number.
```

Note: Take User input

(b) Predict the outcome of the following R commands. **(5M)**

i. print(rep(1:5, times = c(3,1,6,0,2)))
ii. print(list(c(3, 5, 2), c("apple", "banana", "cherry"),c('a','b','c'))[[2]][2])
iii. print(matrix(1:9,ncol =3, byrow = T)[,1])
iv. aa<-c(1,2,'3'); print(aa); class(aa)
v. print(seq(5,50, by = 3))

Q4. Consider the following data:

x = (3, 5, NA, 15, NA, 25, 3, 5)
y = (10, 13, 10, NA, 10, 10, 13, 10)

Write commands using R to implement the following queries and show the corresponding outputs:

i. Create x and y vectors and create a data frame 'DFrame' with x and y vectors. **(4M)**
ii. Add row names to the DFrame using letters(1: length(x)) function and display the DFrame.**(2M)**
iii. Count the number of 'NA' values in both x and y attributes. **(2M)**
iv. Replace all 'NA' values in both x & y attributes with the mean of x& y respectively.**(2M)**

QP MAPPING

Q. No.	Module Number	CO Mapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	1	1	1, 2, 3	1, 2, 3, 4	1, 3	15
Q2	1	1	1, 2, 3	1, 2, 3, 4	1, 3	10
Q3	2	2	1, 2, 3	1, 2, 3, 4	2	15
Q4	2	2	1, 2, 3	1, 2, 3, 4	2	10

VIT-AP
UNIVERSITY
Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2023-2024–CAT-1

Course Code: CSE1006 **Course Title: Foundations for Data Analytics**

Set number: 6 **Date of Exam:** 08/02/2024 (AN) (D2)

Duration: 90 Minutes **Total Marks: 50**

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. (a) Discuss various types of data analytics to improve decision-making.**(5M)**
(b) Citing an example, explain how Data analytics helps in Risk Management. **(5M)**

Q2. (a) Citing an example, differentiate between Data and Information. **(5M)**

(b) Explain below terms related to Data analytics. **(10M)**
i. Data Scraping
ii. Data Munging
iii. Data Sampling

Q3. (a) Write a function in R that accepts an integer value and returns the factorial of it. **(5M)**
(b)Observe the below R commands and predict their outcomes.**(10M)**

```
i.vec<- c(2, 4, 6, 8, 10)
  print(vec * 2)
  print(vec[[2]])

ii.values<- c(3, 5, 2, 7, 4)
   cumulative_sum<- cumsum(values)
    print(cumulative_sum)

iii. x<<-2
    3 ->> z
    print(x^y)
    print(class(x))
    print(typeof(x))

iv. print(5/2)
    print(5%%2)
    print(5%/%2)
```

Q4.Consider the below given ‘Student’ dataset and answer the following questions: **(10M)**

stu.id	stu.name	Age	Address
1	John	24	United States
2	Anne	28	London
3	Paul	20	Paris
4	Joyce	22	Germany

i. Create a data frame to store the 'Student' dataset.
ii. Write a R command to extract 2nd and 3rd records from the student dataset.
iii. Write a R command to display the Age attribute of all students.
iv. Write a R command to add a new record to the student dataset with stu.id 5.
v. Write a R command to add a new attribute Department to the student dataset.

QP MAPPING

Q. No.	Module Number	COMapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	1	1	1, 2, 3	1, 2, 3, 4	1, 3	10
Q2	1	1	1, 2, 3	1, 2, 3, 4	1, 3	15
Q3	2	2	1, 2, 3	1, 2, 3, 4	2	15
Q4	2	2	1, 2, 3	1, 2, 3, 4	2	10

VIT-AP
UNIVERSITY

QUESTION PAPER

Name of the Examination: Winter 2023-24Semester–CAT-1

Course Code: CSE1006 **Course Title: Foundations for Data Analytics**

Set number: 7 **Date of Exam:** 09/02/2024 (AN) (E2)

Duration: 90 Minutes **Total Marks: 50**

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. (a) If the mean of 28, X, 42, 78, and 104 is 62, then what is the mean of 48, 62, 98, 124, and X? **(2M)**

(b) The mean of 5 numbers is 6. The mean of 3 of them is 8. Then, what is the mean of the remaining two numbers?**(2M)**

(c) The average weight of a group of 30 friends increases by 1 kg when the weight of their football coach was added. If average weight of the group after including the weight of the football coach is 31 kg, then what is the weight of their football coach? **(3M)**

(d) The arithmetic mean of the 5 consecutive integers starting with s is a. Then, what is the arithmetic mean of 9 consecutive integers that start with s + 2? **(3M)**

Q2. (a) Consider the following set of data:

108, 99, 112, 111, 108, 98

Normalize it using the following methods:

(i) Decimal Scaling **(2M)**

(ii) Min-Max normalization with a range between [10, 20] **(3M)**

(iii) Z-Score **(5M)**

(b) What do you mean by noise in the data? Explain with an example. **(5M)**

Q3. (a) Define a function that accepts an integer vector and returns the count of those data elements that are both positive and even. **(5M)**

(b) Write an R program to input an integer matrix of dimension 3x3 and print the sum of the principal diagonal elements. **(5M)**

(c)Define a function that accepts an integer matrix of dimension 3x3 and returns a vector consisting of the mean of each row of the matrix. **(5M)**

Q4. Predict the outcome of each of the following print statements. **(5*2=10M)**

```
print(sum((0:5)[-1]))
print(dim(list(list(1:10),list(11:20))))
print(length(seq(1,10,3)))
print(array(1:18,c(3,3,2))[1,,])
print(class(c('T','T','T','T')))
```

QP MAPPING

Q. No.	Module Number	COMapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	1	1	1,2,3	1,2,3,4	1,3	10
Q2	1	1	1,2,3	1,2,3,4	1,3	15
Q3	2	2	1,2,3	1,2,3,4	2	15
Q4	2	2	1,2,3	1,2,3,4	2	10

VIT-AP UNIVERSITY

QUESTION PAPER

Name of the Examination: Winter 2023-24Semester–CAT-1

Course Code: CSE1006

Course Title: Foundations for Data Analytics

Set number: 8

Date of Exam: 10/02/2024 (AN) (F2)

Duration: 90 Minutes

Total Marks: 50

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. Consider the following population of scores and perform the below given operations on it.

88, 93, 103, 118, 113, 101, 101, 97, 115, 99, 102, 97, 100, 107, 116, 107, 110, 113, 95, 89

(i) Simple Random Sampling with Replacement and a sample size of 5 **(2M)**

(ii) Simple Random Sampling without Replacement and a sample size of 5**(2M)**

(iii) Stratified Sampling with four strata and a sample size of 5**(3M)**

(iv) Cluster Sampling with a sample size of 2**(3M)**

Q2. (a) Consider the following set of data:

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 94, 96

Remove noise from the above data using the following techniques::

(i) Equal depth partitioning having bin depth of 3 and smoothing by bin means **(2M)**

(ii) Equal depth partitioning having bin depth of 3 and smoothing by bin boundaries **(2M)**

(iii) Equal depth partitioning having bin depth of 3 and smoothing by bin medians **(2M)**

(iv)Equal width partitioning having width of 10 **(4M)**

(b)What are the major steps involved in data analytics? Briefly, discuss them. **(5M)**

Q3. (a) Write an R program to create an integer vector of length 10 and print the second largest element present in it. **(5M)**

(b)Consider the below given data frame student and write R commands to execute the following tasks:

```
id=1:10
name=letters[1:10]
age=sample(18:22,10,replace=T)
score=sample(1:50,10,replace=T)
student=data.frame(id,name,age,score)
```

(i) Print the range of the score attribute.**(2M)**

(ii) Print the details of those students whose score is more than 30.**(2M)**

(iii) Print the details of those students who have received the maximum score. **(2M)**

(iv) Print the names of those students whose age is less than 20. **(2M)**

(v) Print the identities of those students whose score is more than the mean of all the score values. **(2M)**

Q4. Consider the following array 'arr' and predict the outcome of each of the following print statements.**(5*2=10M)**

```
arr=array(1:9,c(3,3,2))
print(sum(arr[,,1]))
print(sum(arr[,,1]==arr[,,2]))
print(arr[1,,1]==arr[1,,2])
print(max(arr[,1,1]))
print(arr[,1,])
```

QP MAPPING

Q. No.	Module Number	CO Mapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	1	1	1,2,3	1,2,3,4	1,3	10
Q2	1	1	1,2,3	1,2,3,4	1,3	15
Q3	2	2	1,2,3	1,2,3,4	2	15
Q4	2	2	1,2,3	1,2,3,4	2	10

VIT-AP
UNIVERSITY
Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2023-2024–CAT-1

Course Code: CSE1006 **Course Title: Foundations for Data Analytics**

Set number:9 **Date of Exam:** 09/02/2024 (FN) (E1)

Duration: 90 Minutes **Total Marks: 50M**

Instructions:

1. Assume data wherever necessary.
2. Any assumptions made should be clearly stated.

Q1. a) How can the characteristics of data influence its reliability? **(5M)**

b) Define Data and Information. Compare Data Vs Information. **(5M)**

Q2. a) Consider the following table of data and answer the following questions.

Weight (kg)	Height (ft)	Gender
50	6	F
65	5.8	F
70	6.2	M
NA	6.4	M
68	NA	F
NA	5.5	F

Answer the following:

i. Find the average of weight and height **(3M)**
ii. Find and replace the NA with mean of the Weight **(3M)**
iii. Replace Gender Male with '0' and Female with '1' **(4M)**

b) How would you describe the data sampling and data cleaning in Data Analytics? Give some examples. **(5M)**

Q3. Write **R commands** and **outputs** for the following: **(10M)**

i) Create a vector **numbers** by using seq () function from 1 to 50 with an interval of 5.

ii) Extract the elements from index position 3 to 9 in **numbers** vector.

iii) Display items for specified indices 1,3,6,9 in **numbers** vector.

iv) Extract the elements except the first 3 elements in **numbers** vector.

v) Repeat each and every element with 2 times in **numbers** vector

Q4.Consider the below given 'Student' dataset and answer the following questions

Marks1	Marks2	Marks3
79	65	80
82	61	55
96	70	70
54	65	79

i. Create a data frame "Student Marks" and display the data frame. **(3M)**
ii. Write an R command to assign row names Mary, Alice, Bob, Judy respectively. **(3M)**
iii. Write an R command to change column names to Python, FDA and SE respectively. **(3M)**
iv. Write an R command to extract Mary and Judy details. **(3M)**
v. Write an R command to extract FDA marks for all students **(3M)**

QP MAPPING

Q. No.	**Module Number**	**COMapped**	**PO Mapped**	**PEO Mapped**	**PSO Mapped**	**Marks**
Q1	1	1	1, 2, 3	1, 2, 3, 4	1, 3	10
Q2	1	1	1, 2, 3	1, 2, 3, 4	1, 3	15
Q3	2	2	1, 2, 3	1, 2, 3, 4	2	10
Q4	2	2	1, 2, 3	1, 2, 3, 4	2	15