VIT-AP UNIVERSITY

Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2022-2023 – CAT-2

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

Set number: 3 **Date of Exam:** 29/03/2023 (AN)

Duration: 90 mins **Total Marks: 50** (A1)

Instructions:

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

Q1. Create a data frame as depicted in the below table. Write R code for the below operations **(10M)**

Car brand	KM driven	Price in Dollars
Lexus	105.546	4600
Toyota	216.924	2500
Toyota	216.924	2500
Lexus	107.355	5700
Mercedes-Benz	209.466	4700
Lexus	135.819	3500
Mercedes-Benz	120.785	4500
Lexus	156.704	3300
Lexus	162.262	4700
Land Rover	132.352	5000
Toyota	170.599	3500

a. Create the data frame (2M)
b. Sort the data frame based on **KM driven** (2M)
c. Sort the data frame based on **KM driven** and **Price in Dollars** (2M)
d. Sort the **KM driven** in ascending order and **Price in Dollars** in descending order (2M)
e. Find the maximum value in **KM driven** and **Price in Dollars**. (2M)

Q2. Perform operations on the below given data frame. **(10M)**

```
A <- data.frame(
Gender = c("F", "F", "M", "F", "B", "M", "M"),
Height = c(154, 167, 178, NA, 169, 183, 176))
```

a. Fill the **NA** with the mean value. (2M)
b. Recode the **Height** which are greater than 170 to 180. (2M)
c. Recode the **Height** which are greater than 160 and less than 170 to 165. (2M)
d. Recode **F to Female, M to Male** and the remaining to **Others.** (2M)
e. Create one more called **Result** from **Height** column. If Height is greater than 170 Result should be good, else result should be Bad. (2M)

Q3. Given two tables write R code and its output to perform various join operations. **(10M)**

Table1=data.frame(
TrainNo=c(123,234,345,456,678,789),
Source=c("A","B","C","D","E","F"),
Destination=c("U","V","W","X","Y", "X"))

Table2=data.frame(
TrainNo=c(123,234,345,789, 102,103),
NoofPassenger=c(77,89,92,68,109,99))

a. Write the R code and output for **Inner join** (2M)
b. Write the R code and output for **Outer join** (2M)
c. Write the R code and output for **Left Outer join** (2M)
d. Write the R code and output for **Right Outer join** (2M)
e. Write the R code and output for **Cross join** (2M)

Q4. Write R programming code for the operations listed below. **(10M)**

a. Import a text file (example.txt) into the R programming environment. (2M)
b. Write the variable (df) into a comma separated file called **output.csv.** (2M)
c. Write a function to import set of values from a text file to vector. (2M)
d. Draw a bar chart for the data frame **Readings** and the columns are **Months** (x-axis) and **Temperature** (y-axis) (2M)
e. Draw a histogram for the values c (5, 9,12, 13,21,8,31,36,22,12,41,31,33,19). (2M)

Q5. Write R programming code for the operations listed below using dplyr package. **(10M)**

a. Write a function to sample 20% of data from data frame named ***df***. (2M)
b. Remove duplicated rows in the data frame ***df***. (2M)
c. Retrieve all columns from the dataset called **students** and all rows where **total** is greater than 450. (2M)
d. Retrieve all columns from the dataset called students and exclude the rows whose **total** is less than 200. (2M)
e. Find mean and median for the column **total** in the **students** data frame. (2M)

QP MAPPING

Q. No.	Module Number	CO Mapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	3	1	1, 2, 3	1	-	10
Q2	3	1	1, 2, 3	1	-	10
Q3	3	1, 2	1, 2, 3	1, 2	-	10
Q4	4	1, 2	1, 2, 3	1, 2	-	10
Q5	4	1, 2	1, 2, 3	1, 2	-	10

VIT-AP UNIVERSITY

Apply Knowledge. Improve Life!®

QUESTION PAPER

Name of the Examination: WINTER 2022-2023 – CAT-2

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

Set number: 4 **Date of Exam:** 01/4/2023 (AN)

Duration: 90 mins **Total Marks: 50** (A)

Instructions:

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

Q1. Construct a data frame as shown in the table and write R code for below operations. **(10M)**

Col-1	Col-2
A	1
A	1
A	2
B	2
B	3
B	3
C	2
C	3
C	4
D	1
D	1
D	2

a. Remove the duplicated rows and print only the unique rows (2M)
b. Print only the duplicated rows in the data frame (2M)
c. Consider the data frame is stored in a variable called ***df***. What is the output of the snippet? ***duplicated (df,)*** (2M)
d. Display the duplicate records based on the **Col-1** in the data frame. (2M)
e. Remove the duplicate records based on the **Col-2** in the data frame. (2M)

Q2. Perform the operation on the data frame. **(10M)**

```
df <- data.frame(col1 = c(1:3, NA),
col2 = c("this", NA,"is", "text"),
col3 = c(TRUE, FALSE, TRUE, TRUE),
col4 = c(2.5, 4.2, 3.2, NA))
```

a. Write the output of the code ***is.na(df)*** (2M)
b. Write R code to count the total number of NAs in the data frame (2M)

c. Write R code to find the mean and median of **col4** (2M)
d. Count the number of rows without NAs. (2M)
e. Print the index of NAs in the data frame (2M)

Q3. Given two tables write R code and its output to perform various join operations. **(10M)**

Company

CompanyId	CompanyName
1	DELL
2	HP
3	IBM
4	Microsoft

Candidate

Candidat	FullName	CompanyId
1	Ron	1
2	Pete	2
3	Steve	3
4	Steve	NULL
5	Ravi	1
6	Raj	3
7	Kiran	NULL

a. Write the R code and output for **Inner join** (2M)
b. Write the R code and output for **Outer join** (2M)
c. Write the R code and output for **Left Outer join** (2M)
d. Write the R code and output for **Right Outer join** (2M)
e. Write the R code and output for **Cross join** (2M)

Q4. Write R programming code for the operations listed below using dplyr package. **(10M)**

a. Consider a data frame called **student.** Retrieve only the rows where **StudentGrade** is **'A', 'B'** and **'C'** (2M)
b. Consider a data frame called **student.** Retrieve only the rows where **StudentGrade** is **'A', 'B','C'** and **age** is less than 23. (2M)
c. Consider a data frame called **student.** Retrieve only the rows where **StudentName** has the string **'ar'** in it. (2M)
d. Select the column names starts with an alphabet **"S"**. (2M)
e. Exclude all the column names starts with an alphabet **"S"**. (2M)

Q5. Write R programming code for the operations listed below. **(10M)**

a. Plot the pie chart for a data frame **Student** and for the column **Gender** (2M)
b. Plot a scatter plot for a data frame **Employee** and the columns **Age** (x-axis) and **Experience** (y-axis). (3M)
c. Draw a bar chart for the data frame **Readings** and the columns are **Months** (x-axis) and **Temperature** (y-axis) (3M)
d. Draw a histogram for the values c (9,13,21,8,36,22,12,41,31,33,19). (2M)

QP MAPPING

Q. No.	Module Number	CO Mapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	3	1	1, 2, 3	1	-	10
Q2	3	1	1, 2, 3	1	-	10
Q3	3	1, 2	1, 2, 3	1, 2	-	10
Q4	4	1, 2	1, 2, 3	1, 2	-	10
Q5	4	1, 2	1, 2, 3	1, 2	-	10

VIT-AP UNIVERSITY

Apply Knowledge. Improve Life!

QUESTION PAPER

Name of the Examination: WINTER 2022-2023 – CAT-2

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

Set number: 5 | **Date of Exam:** 03/4/2023 (AN) (Cy2)

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

Instructions:

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

Q1. Consider the following code to create a data frame df:

a<-c(5,NA,8,NA,9,10)

b<-c(3,NA,5,23,NA,11)

df<-data.frame(a,b)

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

(i) Count the number of missing values present in ***df***. **(2M)**

(ii) Return the indexes of the missing values present in ***df***. **(2M)**

(iii) Replace all the missing values present in the attribute '**a**' of ***df*** with the mean of the same attribute. **(2M)**

(iv) Replace all the missing values present in the attribute '**b**' of ***df*** with 0. **(2M)**

(v) Replace all the missing values in ***df*** with 'ZERO'. **(2M)**

Q2. Consider the following code to create a data frame ***df***:

x<-c(3,5,10,15,20,25,3,5)

y<-c(10,13,10,10,10,10,13,10)

df<-data.frame(x,y)

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

(i) Sort **df** in the increasing order of both the attributes '**x**' and '**y**'. **(3M)**

(ii) Sort **df** in the decreasing order of both the attributes '**x**' and '**y**'. **(3M)**

(iii) Sort **df** in the increasing order of the attribute '*x*' with the decreasing order of the attribute '**y**'. **(4M)**

Q3. Consider the following code to create two data frames df1 and df2 respectively:

df1<-data.frame(ID=c(1:4),Subject=c('OS','DBMS','FDA','C'))

df2<-data.frame(ID=c(3:6),City=c('Pune','Kochi','Delhi','Trichy'))

Perform inner join, left join, right join, full join, and cross join between them and show the respective outputs. **(10M)**

Q4. Consider the following code to create a data frame **df**:

```
library(dplyr)
id<-c(1,2,3,4,5)
score<-c(78,82,89,94,68)
age<-c(18,19,19,20,21)
gender<-c('M','F','F','M','F')
df<-data.frame(id,score,age,gender)
```

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

(i) Select the attributes containing letter '**e**' in their column names. **(2M)**

(ii) Display the rows where the score is greater than 85. **(2M)**

(iii) Change the variable name from '**score**' to '**marks**'. **(2M)**

(iv) Display the rows either the gender is '**F**' or age is **20**. **(2M)**

(v) Display the mean score of all the rows where the gender is '**F**'. **(2M)**

Q5. a. Consider the below given vector 'age':

```
age <- c(25,26,27,26,28,29,28,30,31,30)
```

Write an R code to create a histogram on 'age'. Properly specify the title of the plot, labels for the axes, colour for the bars, and colour for the border. **(5M)**

b. Consider the following R code to create a data frame 'df':

```
person=c('A','B','C','D','E','F','G')
score=c(2.87,3.41,3.43,1.67,0.83,2.85,1.34)
df<-data.frame(person,score)
```

Write an R code to create a bar plot using 'person' as X-axis and 'score' as Y-axis. Properly specify the title of the plot, labels for the axes, colour for the bars, and colour for the border. **(5M)**

QP MAPPING

Q. No.	Module Number	CO Mapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	3	3	1,2,3	1,2,3,4	-	10
Q2	3	3	1,2,3	1,2,3,4	-	10
Q3	3	3	1,2,3	1,2,3,4	-	10
Q4	4	3	1,2,3	1,2,3,4	-	10
Q5	4	3	1,2,3	1,2,3,4	-	10

VIT-AP UNIVERSITY

Apply Knowledge. Improve Life!®

QUESTION PAPER

Name of the Examination: WINTER 2022-2023–CAT-2

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

Set number: 1 **Date of Exam:** 29/03/2023 (AN)

Duration:90 min **Total Marks: 50** (CC2)

Instructions:

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

Q1. Consider the following data frame and perform the specified operations using R commands to clean the data. **(10 marks)**

Player id	Number of matches	Runs in test series	Runs in one day matches	Highest score
P01	27	196	46	46
P02	NA	324	23	67
P03	32	NA	NA	38
P04	NA	NA	12	15
P05	12	123	8	NA
P01	27	196	46	46

a. Identify the duplicate entry. (2M)
b. Remove the duplicate entry in the data frame. (2M)
c. Replace the NA in **Number of matches** with a constant value 100. (2M)
d. Replace the NA in **Runs in test series** and **Runs in one day matches** with the mean value. (2M)
e. Replace the NA in **Highest score** with "Zero".(2M)

Q2. Consider the following data frame and perform the specified operation using R. **(10 marks)**

Player id	Number of matches	Runs in test series	Runs in one day matches	Highest score
P01	27	196	46	46
P02	52	324	23	67
P03	32	235	38	38
P04	11	12	12	15
P05	12	123	8	9
P01	27	196	46	46

a. Order the data frame according to the **Number of matches** (2M)
b. Sort the data frame in ascending order of **Number of matches** (2M)
c. Sort the data frame in the descending order of the **Highest score.** (2M)

d. Sort the data frame with **Runs in the test series** in increasing order and **Runs in the one day matches** with decreased order. (4M)

Q3. Consider the below given data frames and perform the following operations **(10 marks)**

a. inner join (2M),
b. left outer join (2M),
c. right outer join (2M),
d. cross join (2M)
e. full join (2M).

Data frame 1:

Roll no	Student name	Class	Address
1	Adhi	10^{th}	AA
2	Abi	11^{th}	BB
3	Anu	12^{th}	CC
4	Asha	10^{th}	DD
5	Akshay	11^{th}	EE
6	Aparna	12^{th}	FF

Data Frame 2:

Roll no	Ambition	Specialization
1	Doctor	Biology
2	Engineer	Computer
3	IAS	Social studies
4	Doctor	Biology
5	TravelExplorer	Geography
6	Journalist	Social studies

Q4. Consider the following data frame and write R code to perform the specified operation using dplyr package. **(10 marks)**

EMP_ID	EMP_NAME	CITY	SALARY	AGE
1	Angelina	Chicago	200000	30
2	Robert	Austin	300000	26
3	Christian	Denver	100000	42
4	Kristen	Washington	500000	29
5	Russell	Los angels	200000	36
6	Marry	Canada	600000	48

a. Extract random 3 rows of the data frame. (2M)
b. Select the column that starts with "CITY" . (2M)
c. Select the data in the column that contains "rr" in EMP_NAME. (2M)
d. Print first 3 columns of the data frame. (2M)
e. Summarize the data frame. (2M)

Q5. i. Write an R code to draw the bar chart of the following data with specified properties. **(10 marks)**

Country=c('India', 'Australia', 'Pakistan', 'New Zealand', 'South Africa', 'Sri Lanka')
Matches_Won=c(76,54,32,62,48,23)

a. With proper x and y label (1M)
b. With title of the plot (1M)
c. Specify color for the bars (1M)
d. Specify border width (1M)
e. Limit for x and y axis (1M)

Q5.ii. A pharmaceutical company manufactures 2 different compositions of medicines for a disease. The sale for 6 months is shown in the table. Represent it using a bar chart .

Month	Tablets Sold	
	Composition 1	Composition 2
January	4500	1600
February	2870	5645
March	3985	8900
April	6855	8976
May	3200	5678
June	3456	4555

a. With proper x and y label (1M)
b. With title of the plot (1M)
c. Specify color for the bars (1M)
d. Specify border width (1M)
e. Limit for x and y axis (1M)

QP MAPPING

Q. No.	Module Number	COMapped	PO Mapped	PEO Mapped	PSO Mapped	Marks
Q1	3	3	1,2,3	1,2,3,4	-	10
Q2	3	3	1,2,3	1,2,3,4	-	10
Q3	3	3	1,2,3	1,2,3,4	-	10
Q4	4	3	1,2,3	1,2,3,4	-	10
Q5	4	3	1,2,3	1,2,3,4	-	10