Unique Paper Code	:	32347608
Name of the Course	:	B.Sc. (Hons.) Computer Science
Name of the paper	:	Introduction to Data Sciences
Semester	:	VI
Duration	:	3 Hours.
Maximum Marks	:	75
For admissions of year	:	2015, 2016, 2017 & 2018

Instructions to the Candidates
Attempt any FOUR questions. All questions carry equal marks.

Q1. Write a script containing a function *func* that takes a numeric vector *v* of length 15 as a parameter. Create a dataframe *df* in the function with a column *C1* having values as vector *v*. Call another function *newfunc* to add a column *C2* to *df*. Populate it with *E* or *O* if the data in column *C1* is even or odd, and add another column *C3* to *df* containing random small letters ranging from d to i. Now, print the values of column C1 only, grouped with respect to the column C3. Write a statement in *func* to store the maximum of each column of *df* in a list *l*. Append the output in list *l* as a new row in *df*. Write another statement in *func* to print the summary of column C1 based on column *C3*. Finally return the dataframe *df* as the output. Redirect the output of the script to a file *ff.txt* rather than printing on console.

Q2. The data stored in the array *arr* is shown below.

```
, , Department = D1
                    Role
Gender          Student   Faculty
    Male           15         4
    Female         23         3


, , Department = D2
                    Role
Gender          Student   Faculty
    Male           25        14
    Female         13        10


, , Department = D3
                    Role
Gender          Student   Faculty
    Male            5         1
    Female         24         6
```

Give the output of the following R script and explain the working of each statement:

```
ftable(arr)
order(arr)
as.data.frame(arr)
apply(arr, 2, max)
which(arr==min(arr))
arr[, 2, ] – 1
dimnames(arr)
```

Q3. The description of the dataset “student.csv” is given below.

Variable	Description
Roll	Unique Identification no. of the student.
Name	Name of students
Sex	Factor with levels Male, Female
marks1, marks2, marks3 and marks4	Marks of 4 subjects are given

Some of the values in different marks variable are marked with **NA** (missing values). Create an R script to do the following.

Read student.csv file into a dataframe ‘student’. Find the average marks for each subject without considering missing values. Replace the missing values for each subject with its corresponding average value. Count the total number of students whose marks3 is greater than marks4. Add a new variable(column) ‘total_marks’ in the ‘student’ dataframe that takes the value as sum of all four marks. Create a vector ‘m’ containing roll number of students securing total_marks greater than the average of total_marks. Create a bar chart (with appropriate labels) of total_marks for the top 20 students.

Q4. Write an R script to do the following.

Create a sample of 50 random numbers from a normal distribution with mean 40 and standard deviation 10 and store the result in a vector ‘v1’. Create another sample of 50 random numbers from a normal distribution with mean 42 and standard deviation 9 and store the result in a vector ‘v2’. Create a scatter plot between v1(X-axis) and v2 (Y-axis) and draw two lines of which one in red colour parallel to x-axis and another in green colour parallel to y-axis at their mean value points. Which test is used to compare the mean between two group of samples? Write a command to perform statistic test to compare the means of two groups (v1 & v2).

Let the result of the above test statistic be as follows.

```
data:  x and y
t = -1.2218, df = 97.071, p-value = 0.2248
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.893569  1.402308
sample estimates:
```

```
mean of x mean of y
 40.47095  42.71658
```

Explain the meaning of each variable and provide the interpretation of the result.

Q5. Create R markdown file with the title “*Student Information*”, which contains three chunks:

- First chunk is named “*setup*”. This should set default values for hiding code and results for all chunks.
- Second chunk is named “student_*info*” containing a dataframe with the following columns: *Stud_id*, *Height*, *Weight*, *Age*. Insert 5 values in this field. This chunk should only display dataframe as html tabular output.
- Third chunk is named “plot” and it should show code and display bar chart of *Age* with student id as the labels of the bars. Set the values of height and width of the figure to 6.

Give the description of each chunk in italics and chunk name as H1 heading. Write the command to execute the markdown file as pdf and word output.

Q6. Consider the following three csv files separated by semicolon with header names as:

a. Student_info.csv (roll_no, name, dob, address, phone_no)
b. Paper_Details.csv (paper_code, paper_name)
c. Academic_details.csv (roll_no, paper_code, attendance, percentage).

Write an R script to load packages to connect with the MySQL database server with password credentials. Read all three csv files and load as tables in the database *student* and overwrite if already existing. Write the command to list the tables in the database. Write SQL queries in R to list names of students and names of papers where students have failed (percentage < 40%) and attendance is less than 60% in paper with paper code “*CS7*”. List names of papers with count of students having attendance less than 67%. Find the total students for each paper and visualize top 5 papers using a suitable plot. Finally disconnect the connection from the database server.