Roll Number:____________________

Thapar Institute of Engineering and Technology Patiala
Computer Science and Engineering Department
EST-2022

BE Third Year (5th Semester) 5 Dec. 2022 **UCS548: Foundations of Data Science**
Time: 9.00 AM (3 Hours) **Max Marks: 40**

Instructors: Dr. Sharad Saxena, Dr. Geeta Kasana, Dr. Seema Wazarkar

Note: Attempt all questions. All parts of a question must be answered in order. Assume any missing data.

Q1. a) Consider the sample dataset "**sample.xls**" having 22 records as reference having repeated values of **Area_Code** and **Sales_Code** and is given in Table 1. 5

Table 1: Sample data set

S.No.	YEAR	PRODUCTION (Units)	CONSUMPTION (Units)	Area_Code	Sales_Code
1	2022	23454	15345	4553	44567
2	2021	23343	1456	3334	45565
3	2020	11223	3345	2232	44567
:	:	:	:	:	:
22	2000	14223	13345	2232	44067

Write the R script(s) including *Mapper()* and *Reducer()* concept to perform the following operations.

i) Print *minimum* CONSUMPTION that has lowest occurrence of *Area_Code*. (2.5 marks)
ii) Print the YEAR and PRODUCTION that has highest occurrence of *Sales_Code*. (2.5 marks)

Q1. b) Draw the General Hadoop Architecture, and discuss its various components in detail. 3

Q2. a) Write the HDFS command to perform the following operations: (0.5 marks each) 2

i) Create a directory named INPUT at location /user/.
ii) Copy a file ***name.txt*** into the INPUT directory from ***local*** file system to ***hdfs*** file system.
iii) Display the content of the file ***name.txt***
iv) Remove the file ***name.txt*** from INPUT directory.

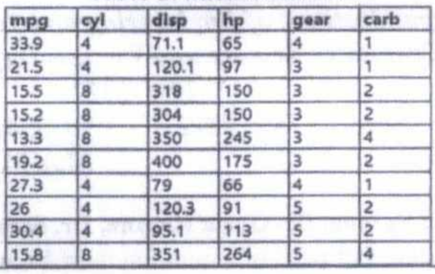
Q2. b) Write R script to create a matrix of size (6 × 10) having random integers values between 0 to 11and perform the following operations on it. (2 marks each) 6

i) Print the count of entries in **each row** which are greater than 4.
ii) Print the **rows** which contain **exactly two** occurrences of the number 7.
iii) Print those **pairs of columns,** whose total (over both columns) is greater than 55. The answer should be a matrix having two columns. For example, the columns 1 and 2 in the output matrix mean that the sum of columns 1 and 2 in the original matrix is greater than 55. Repeating a column is permitted; so, for example, the final **output matrix** could contain the columns (1 and 2), (2 and 4), (2 and 3) etc.

Q3. a) Use *ggplot2* to answer the following questions- 5

i) Write R script to fit a linear model over scatter plot. (2 marks)
ii) State the role of ***facet_wrap()*** and ***facet_grid()***. (1 marks)
iii) Consider ***mtcars*** database shown below (Table 2) with features (***carb, mpg, cyl, disp, hp, gear***). Utilize ***gear*** attribute value in the range (3, 4, 5), and write R statement to draw bar chart with title "Cars with number of Gears". Label the X axis with "Number of Gears" and Y axis with "Count of cars" and bar color should be "blue". (2 marks)

Table 2: *mtcars* dataset



mpg	cyl	disp	hp	gear	carb
33.9	4	71.1	65	4	1
21.5	4	120.1	97	3	1
15.5	8	318	150	3	2
15.2	8	304	150	3	2
13.3	8	350	245	3	4
19.2	8	400	175	3	2
27.3	4	79	66	4	1
26	4	120.3	91	5	2
30.4	4	95.1	113	5	2
15.8	8	351	264	5	4

Q3. b) Consider the following dataset. 3

Ages (year)	1	3	5	6	7	8	10	11	12	14
No. of Kids	1	2	4	2	2	5	4	2	3	3

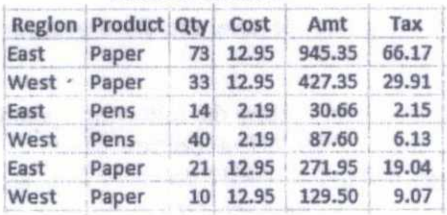
Write the R statement for the follwing:-

i) Show case the use of ***seed()*** to create two groups G1 and G2 having 5 random "**Ages (year)**" with replacement. (1 marks)
ii) Print distinct ages considered in G1, and G2. (0.5 marks)
iii) Print "**No. of Kids**" available in G1, and G2 identified in (2) above. (1 marks)
iv) Write single line R code/function to display the occurrence of each age (**Ages (year)**) in G1. (0.5 marks)

Q4. a) State the importance of Principal Component Analysis (PCA) before data clustering. 2

Q4. b) Consider the data given in Table 3. 6

Table 3: Dataset



Region	Product	Qty	Cost	Amt	Tax
East	Paper	73	12.95	945.35	66.17
West	Paper	33	12.95	427.35	29.91
East	Pens	14	2.19	30.66	2.15
West	Pens	40	2.19	87.60	6.13
East	Paper	21	12.95	271.95	19.04
West	Paper	10	12.95	129.50	9.07

Write R code to accomplish the given task.

i) Create two files with the name "Sales.txt" and "Region.txt", where **Sales** file have Product, Qty, Cost, Amt, Tax data and **Region** file will have data about Product and Region. (1 marks)
ii) Using "Sales.txt" compute **covariance** between Cost and Tax using **pearson** method. (1 marks)
iii) Use "Sales.txt" and "Region.txt" files to display total tax in the East region for Paper. (1.5 marks)
iv) Use "Sales.txt" and "Region.txt" files to display **region name** where maximum number of **pens** sale out. (1.5 marks)
v) Write the R functions to rename "Sales.txt" as "Sales22.txt" and "Region.txt" as "Region22.txt" and to check if "Sales.txt" and "Region.txt" file still exists or not. (1 marks)

Q5. a) State the characteristics and applications of SVD. 2

Q5. b) Find the Singular Value Decomposition (SVD) of the given (2x3) matrix **A** and show all the three components U matrix, Σ matrix, and V matrix. 6

(2 marks each)

$$A = \begin{bmatrix} 3 & 2 & 2 \\ 2 & 3 & -2 \end{bmatrix}$$