Ramakrishna Mission Vivekananda Educational and Research Institute
PO Belur Math, Howrah, West Bengal 711 202
School of Mathematical Sciences
Department of Computer Science

MSc Big Data Analytics : Batch 2023-25, Semester 1, End Term
Programming for Data Science
Instructor: Dr. Sudeep Mallick

*S*tudent Name (in block letters): Date: 8 Dec 2023
*S*tudent Roll No: *M*ax Marks: 100
*S*ignature: Time: 3 hrs

Please note the following:

- **Question 1 is COMPULSORY. Attempt ANY TWO other questions. Questions 2,3 and 4 carry equal marks.**
- You are allowed to refer to your class notes, material, etc.
- You are **NOT** allowed use of the Internet during the examination unless instructed otherwise by the invigilator.
- Any indication of plagiarism or infraction would result in either cancellation of the exam or a particular question as the case may be.
- You are STRONGLY advised not to copy code from class examples, you may refer to them however. Any indication otherwise would result in either cancellation of the attempt or suitable deduction of marks as deemed appropriate during evaluation.

1. Attempt the following (COMPULSORY - *all parts carry equal marks*): [8x5 = 40]
   (a) Create a list of marks between 0 and 100 for 500 students. Write a function that accepts a single student's marks and returns his grade (You can use any appropriate grade structure of your choice - A,B,C,D,E). Now use list comprehension to generate a list of grades from the list of marks using the function you just created. How many received A grade? How many received B grade? Hint: you may use count function of list.
   (b) Build a function which accepts a list and 2 numbers and returns the count of the number of times these numbers are found in the given list as two outputs from the function. Provide your test code.
   (c) Build a function which accepts a variable number of parameters - one of the parameters is a number which is going to be used as a check for the rest of the parameters (let's call this parameter as check number). The function checks if the rest of the parameters is divisible by that check number and returns the sum of those divisible numbers. Test with different check numbers and different number of parameters.
   (d) Build a function that accepts filename as the parameter and returns the file contents. Decorate this function with a decorator which computes the time taken to read the file and prints the file name and the time taken. Demonstrate the use of the decorated function. The test code should print the returned file contents.
   (e) Create a matrix of random numbers from 0 to 6 of size 4x5 and retrieve only 2nd and 4th column and save it in a 4x2 matrix. Similarly retrieve only the 1st and 4th rows and save it in a 2x5 matrix.
   (f) Part A. Read the provided titanic data-set and retrieve only the even positioned columns i.e. columns 0,2,4,... without manually providing the column names. Part B. Retrieve just the passenger names who have survived (value 1 in the Survived column indicates that the passenger has survived).
   (g) Create a matrix of numbers from 0 to 6 of size 4x5 and find the row and column sums.

(h) From numpy linear algebra package using only the inv function (and any base numpy module functions) solve the following simultaneous equation

```
2x + y = 5
4x - 3y = 2
```

where A is the matrix of the LHS coefficients and B is the RHS constants so the solution vector is matrix product of the inverse of A and the orginal matrix B. Verify the solution is correct by again doing multiplication of appropriate matrices.

2. Attempt the following questions: [10+10+5+5]

(a) Using numpy arrays draw two sectors arranged vertically joined in the middle as shown in the figure. Your plot should resemble the provided figure as much as possible.

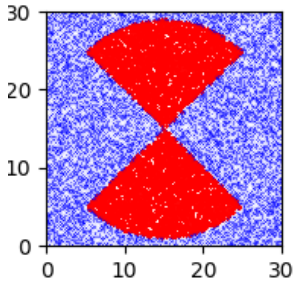

Figure 1: Sectors joined in the middle

(b) Create a base class (parent class / superclass) which captures information about library items such books, magazines, newspapers, etc. It has a constructor which takes title, publisher details and publication date as data and has methods for checking out and returning the item.

Create a class which is a subclass of the above class and captures information about books. The book item has an additional attribute such as authors and genre (type). It also overrides the check out method and introduces a new method which returns information about the genre of the book.

The classes should maintain the checked out status information of the item. You should not be able to check out an already checked out item. Also the return method should reset the checked out status so that the item is eligible for check out. Similarly an item which has not been checked out cannot be returned. Any attempt to return an item not checkout will result in an appropriate error message.

Code the class and sub class in a python script (Not in IPython notebook) and the script should have a separate main block for testing the classes. Include your test code.

(c) Attempt the following parts:

i. Create a function which receives any matrix and returns a transpose of it without using the T property of a numpy array.

ii. Create a function which receives any matrix and returns a sub-matrix of selection of rows where row average value is more than the entire matrix average.

3. Attempt the following questions: [10+2+2+1+1+1+3+2+1+3+1+1+2]

(a) Draw an thick band ellipse where the general equation of ellipse is given below with the a and b as half the major and minor axes of the ellipse centered around x0,y0

$$((x - x_0)/a)^2 + ((y - y_0)/b)^2 = 1$$

as shown in the figure. Your plot should resemble the provided figure as much as possible.

(b) Using the titanic data-set provided attempt the following:

i. Find the distribution (frequency count) of missing values row wise i.e. how many rows have nil missing value, how many have 1 missing value per row, and so on. No manual counting should be required to arrive at the answer.

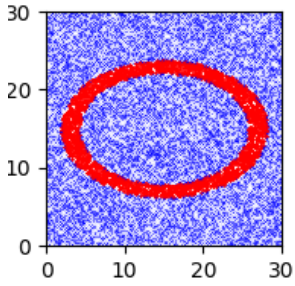

Figure 2: Elliptical band

ii. Find the distribution of missing values column wise i.e. how many columns have nil missing value, how many have 1 and so on. No manual counting should be required to arrive at the answer.

iii. Which column has the maximum missing values. Explain your answer.

iv. Drop the top two columns which have maximum missing values.

v. Has the "missingness" situation improved. Please comment briefly, justify your answer with programmatic outputs.

vi. Ensure that all the int columns are int16 and float columns are float16. Hint: use either of the two approaches - A. re-read the data-set with correct data types and then re-apply the transformation of the above steps or B. directly apply data type changes to the data-set transformed by the above steps.

vii. Whatever missing cell(s) are remaining try to fill it with an appropriate value - column mean value in case of numeric and column median value in case discrete data column.

viii. Find the average fare paid by the survived and those by non-survived. Comment / interpret the result.

ix. Use three different methods to find the average fare paid by the surviving females and those by non-surviving females. Comment / interpret the result.

x. Find the number survived based on combination of sex and class values. Comment / interpret the result.

xi. Find the percentage (or fraction) survived based on combination of sex and class values. Comment / interpret the result.

xii. Reorder columns such that the Name column is the first column

4. Attempt the following questions: [10+12+8]

(a) Using numpy arrays draw bands slanted at 45% separated at some distances as shown in the figure. Your plot should resemble the provided figure as much as possible.

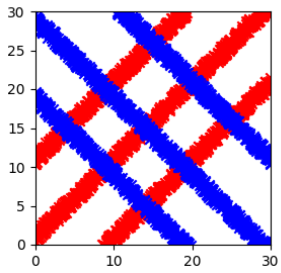

Figure 3: Two coloured bands

(b) Create a function which finds the determinant of a 2x2 matrix, use it to define a function which finds the determinant of a 3x3 matrix. Use this function to find the determinant of a 4x4 matrix. Use 3

test cases of 2x2, 3x3 and 4x4 matrices to demonstrate the correctness of these three functions (to demonstrate correctness use the det function of the numpy linear algebra package).

The three functions you build should not use the numpy linear algebra package. You are to use the linear algebra package only to demonstrate the correctness of the function.

Incorporate exception handling in case non-square matrices or matrices not of type int16 or int32 passed a parameters in all the three functions. Hint: you may offload the exception handling into a fourth function which can be reused by the 3 determinant functions.

(c) Given a string create a dict which shows the elements as keys and the number of occurrences as values. Plot the frequencies of each character as a simple bar graph using matplotlib pyplot's bar method. The string to be used is 'given a string create a dict which shows the elements as keys and the number of occurrences as values'.