Unique Paper Code	:	32347507
Name of the Course	:	B.Sc. (Hons.) Computer Science
Name of the paper	:	BHCS15A Data Analysis and Visualisation
Semester	:	V sem
For admissions of year:		2019 onwards

Duration : 3 Hours Maximum Marks : 75

Instructions to the Candidates:

Attempt any FOUR questions.
All questions carry equal marks.
Answer parts of a question together.
Use Python statements and include appropriate libraries wherever required.

Q 1 Given the following list of strings

List1 = ['Good Morning', 'Good Evening', 'Hello', 'Good afternoon', 'Greetings', 'Good Morning', 'Nice to see you']

Perform the following operations

a) Use list comprehension to store unique strings with multi-words of 'List1' to another list named 'Newlist'. Also, write an anonymous function to sort the given list 'List1' on the last character of each string.

b) Using 'List1', generate the following dictionary 'Anydict' where key is the count of words in a string and value is the list of strings having that count.

Anydict={1:['Greetings','Hello'], 2: ['Good Morning', ' Good Evening', 'Good afternoon', 'Good Morning'], 4: ['Nice to see you']}.

Create a data series 'Ds1' using the created dictionary 'Anydict'.

c) Draw a bar plot to compare the frequency of strings with equal word counts in 'List1'. For example, the frequency of strings with two words in the given list 'List1' is 4. Give proper names to both axes.

Q 2 Consider the following DataFrame **EXERCISE** to answer the given questions where 'Kind' attribute indicates the type of exercise regime followed.

ID	Name	Diet	Pulse	Time (min)	Kind
0	A	low fat	85	40	walking
1	A	low fat	85	45	walking
2	A	no fat	88	30	running
3	B	no fat	90	10	walking
4	B	no fat	92	15	rest
5	B	low fat	93	30	rest
6	C	low fat	97	15	rest
7	C	low fat	97	15	rest
8	C	low fat	94	30	walking
9	D	low fat	80	10	walking
10	D	low fat	82	15	rest
11	D	low fat	83	30	rest
12	E	no fat	91	10	rest
13	E	low fat	92	15	running
14	E	low fat	91	30	running

a) Create a new DataFrame **SELECTED** having a hierarchical index on columns "Name" and "Diet". Then, find the maximum pulse rate for each individual in the SELECTED DataFrame.

b) Count the total number of records of individuals having names 'A' or 'B' and who are following a low fat diet plan from the data frame **SELECTED** created in part (a). Also, sort DataFrame **SELECTED** on index at first level in descending order.

c) Using DataFrame **EXERCISE**, create a figure with two subplots and save the figure with the name 'exerciseplot.jpeg'. Set title of the figure as 'EXERCISE'. First subplot compares the average pulse rate of individuals and the second subplot shows the relationship between variables 'Pulse' and 'Time'. Do color encoding using variable 'kind' in the scatter plot.

Q 3 a) Given the following commands to create series sr

```
import numpy as np
import pandas as pd

sr = pd.Series(['Madhuri','AjaySh@rma', 'R@ni', 'Radha',np.nan,'Smita','3567'])
```

Write separate commands to compute the length of each string in the series, replace @ with 'a' in all strings in the series, count the occurrences of 'a' in each string, change the case of all letters, find all strings with pattern 'adh' in them and find all strings that end with letter 'i'.

b) Create a DataFrame of *7* rows and *7* columns containing random integers in the range of *1* to *100*. Compute the correlation of each row with the preceding row.

c) Write Numpy code to generate a random list of *100* integers (range of *55* to *150*) and identify the index of the largest element and smallest element. Change this list into a *10 x 10* matrix and replace all diagonal elements with 1.

Q4 Using the data frame **EXERCISE** provided in Q*2* , attempt the following questions

a) What is a map function? Use map function to convert all values in the 'Diet' attribute to uppercase.
b) Assuming the data is stored in a csv file "*Exercise.csv"*, give appropriate commands to read this file, indexed on 'Name' and 'Diet' into a dataframe named **EXERCISE**. Modify this command to read only the first 5 rows of the file. If the file contains millions of records then give the command to read the file in small pieces of uniform size.
c) Differentiate between *qcut* and *cut* methods. Use the appropriate method to create *4* bins on the 'Pulse' attribute. Store the corresponding bin value of 'Pulse' attribute as a new attribute 'Pbin' in the original DataFrame. Display the count of values of each bin.

Q5 Consider the following DataFrame **ADM** containing data of freshly admitted students in a college during various rounds of admission. The DataFrame consists of the student's name, cut off list in which he/she has taken admission, date of admission, his/her % of marks, course code and gender.

Sid	Name	List	DateAdm	Marks %	Course Code	Gender
S1	Amit Jaiswal	I	01-07-2021	97	C001	Male
S2	Pradeep Dubey	II	09-07-2021	95	C009	Male
S3	Rinky Arora	I	04-07-2021	90	C112	Female
S4	Sonia Shah	IV	01-08-2021	96	C001	Female
S5	Sushil Negi	III	20-07-2021	96.5	C001	Male
S6	Neeraj Gaur	II	11-07-2021	94.5	C009	Male
S7	Preeti Sharma	IV	03-08-21	89	C112	Female
S8	Deep Gupta	III	23-07-2021	95.75	C001	Male
S9	Priya Bansal	II	10-7-2021	93.5	C009	Female
S10	Anand Ahuja	I	01-07-2021	88.5	C112	Male

Perform the following:

a) Set the first column 'Sid' as the row index of the given DataFrame **ADM**. Create a pivot table of the DataFrame to display the total number of admissions as per 'Course Code' and 'Gender'.

b) For each 'List', find the total number of admissions, minimum 'Marks%' and maximum 'Marks%' in each course.

c) Calculate and display the average 'Marks%' of all Female students of course 'C112'.

Q6 a) Give Pandas statements to create two data series of random floating-point numbers where the first data series has a datetime index of all second Tuesdays of every month of 2021 and the second data series has a datetime index of 20 continuous dates ending at 31/01/2021.

b) What is resampling? Write python code depicting the usage of resample method.

c) Create a DataFrame **DS** with two columns 'Dates' and 'Sale' containing all dates of January 2021 and 31 random integers between 500 and 1000 respectively. Add another column 'Moving Avg' to **DS** containing the rolling average of 5 consecutive values in the 'Sale' column. Plot simple line plots between 'Dates' and 'Sale' as well as 'Dates' and 'Moving Avg'. Explain the utility of the rolling method with respect to these plots.