Unique Paper Code : **42343307**

Name of the Course : **B. Sc. Programme / B. Sc. Mathematical Science SEC-1**

Name of the Paper : **Data Analysis using Python Programming**

Semester : **III**

Year of Admission : **2019 onwards**

Duration: 3 Hours **Maximum Marks: 75**

Attempt any **four** questions.

All questions carry equal marks.

1. Consider a list of values:

```
bag = [25,26,21,22,31,29,33,34,26,30,31,46]
```

- Import the appropriate Python libraries to create a ndarray called `bag_weights` having 3 rows and 4 columns from the list `bag`.
- Use Numpy library to display the mean, variance and median of the given data in `bag_weights`.
- Write a command to display the count of values greater than the median in `bag_weights`.
- Transpose `bag_weights` and then split it in two arrays `bagA` and `bagB` having 2 rows and three columns each.
- Sort `bagA` such that it brings the highest value of the row in the first column. Sort `bagB` such that it brings the lowest value of the row in the first column.
- Find the union and intersection of values in `bagA` and `bagB`.

2. Consider a list of values:

```
rate = [4.23,3.8,2.98,2.56,3,114,3.8,3.78,2.98,4.8,4.10,3.65]
```

- Import the appropriate Python libraries to create a one-dimensional ndarray called `growth_rate` from the list `rate`. Create another one-dimensional array named `twos` having the same number of elements as `growth_rate`, all set to 2.
- Use Numpy library to find the index of the maximum and the minimum values in the array `growth_rate`.

- What does a box plot show? Give a command to display a boxplot for `growth_rate`.
- Concatenate the two arrays `growth_rate` and `twos`, and reshape the resulting array to have four rows and appropriate number of columns, call it `results`.
- Find the mean, median, mode and standard deviation of each column in `results`.
- Write a command to store the array `results` to a file called `result.npy` on the disk in the current working directory.

3. Consider the following DataFrame (`df`):

#	Movie_ title	Director_ name	Language	Length	Budget	Gross_ collections	User_ rating	Critic_ rating
1	AAA	Ram	Urdu	120	90	80	4	7
2	BBB	Eash	Hindi	NULL	65	70	6	6
3	CCC	Anju	Hindi	125	100	150	9	8
4	DDD	Jay	Hindi	150	85	85	6	5
5	EEE	Eash	Hindi	90	60	NULL	7	5
6	FFF	Suraj	French	100	115	120	8	6
7	GGG	Anju	French	NULL	80	81	5	5
8	HHH	Ram	French	115	50	40	3	4
9	JJJ	Anju	French	120	92	75	3	6

Write suitable Python command(s) in Pandas library:

- Display the number of rows and columns present in the DataFrame `df`?
- Display the names of columns that have `NULL` values present in them, along with the count of `NULL` values. Replace the `NULL` values present in the column with the lowest value in that column.
- Create a new column in `df` named `Rating`, which contains the mean of `User_ rating` and `Critic_rating`. Create another column, `Profit`, which contains the difference of `Gross_collections` and `Budget`.
- Find the correlation between `Budget` and `Rating`. Based on the correlation values between two variables, what inference(s) can be drawn about the relationship between them?
- Group the movies according to the `Director_name`. Find the most profitable director.
- What does a contingency table depict? Write commands to display the contingency table between `Director_name` and `Language`.

Q4. Consider a dictionary:

```
dict1 = {Chhetri: 80, Shabbir: 23, Gouramangi: 6,
     Subrata: 92, Vijayan: 29, Gawli: NULL, Nabi: 7,
     Renedy: 4, Lalpekhlua: 23, Baichung:41, Surkumar: 2}
```

Write suitable Python command(s) in Pandas library:

- Create a Pandas Series for the dictionary `dict1` where the key is name of the footballer and the value is the number of goals scored by him. The Series should have the names of the footballers as its index and values as goals scored.
- Display the names of Footballers who have scored more than 20 goals.
- Due to the good performance of top six footballers, their rankings have increased and the number of goals scored by them need to be increased by 25. Round the resulting value to the nearest integer equal to or more than the computed number of goals. Update the Series to reflect these changes.
- Include a 12th man named 'Mondal' in the above Series whose number of goals scored is not known.
- Display the list of Footballers whose number of goals scored is NOT NULL.
- Due to injury, 'Shabbir' was replaced by 'Sandhu' who number of goals scored is 5. Reflect this change in the Series and display the new Series.

Q5. The first few rows of the standard `iris` dataset in the sklearn library are given below:

Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	4.7	3.2	1.3	0.2	Iris-setosa
4	4.6	3.1	1.5	0.2	Iris-setosa
5	5	3.6	1.4	0.2	Iris-setosa
6	5.4	3.9	1.7	0.4	Iris-setosa
7	4.6	3.4	1.4	0.3	Iris-setosa
8	5	3.4	1.5	0.2	Iris-setosa
9	4.4	2.9	1.4	0.2	Iris-setosa
10	4.9	3.1	1.5	0.1	Iris-setosa
11	5.4	3.7	1.5	0.2	Iris-setosa
12	4.8	3.4	1.6	0.2	Iris-setosa
13	4.8	3	1.4	0.1	Iris-setosa
14	4.3	3	1.1	0.1	Iris-setosa
15	5.8	4	1.2	0.2	Iris-setosa
16	5.7	4.4	1.5	0.4	Iris-setosa
17	5.4	3.9	1.3	0.4	Iris-setosa
18	5.1	3.5	1.4	0.3	Iris-setosa
19	5.7	3.8	1.7	0.3	Iris-setosa

- Import the appropriate Python libraries to load the dataset. Create a Pandas DataFrame named `iris` having all the columns in the dataset .

- Use an appropriate command to display a summary of the vital statistics of all numerical and categorical attributes in `iris`.

- What is the role of pre-processing in data analysis? Discuss how will you choose between (a) deleting the rows containing missing values or (b) replacing the missing values in a column with the mean or (c) replacing them with the mode of the column.

- Give a Pandas command to convert the categorical attribute, `species` into dummy variables. Display all the columns of the DataFrame including the dummy variables. Give a command to drop the column `species` from the DataFrame.

- Draw a scatterplot between the columns sepal length and petal length for the species `setosa` in `iris`.

- Create 5 equal length bins for each of the two columns sepal length and sepal width. Draw two histograms, one each for the values of sepal length and sepal width in these bins in a single figure. Save this image in a file on the hard disk.

Q6. Consider the details 15 rubies as follows:

Pc No	carat	cut	color	clarity	depth	table	Price in (thousand INR)
1	0.23	Ideal	E	SI2	61.5	55	326
2	0.21	Premium	E	SI1	59.8	61	326
3	0.23	Good	E	VS1	56.9	65	327
4	0.29	Premium	I	VS2	62.4	58	nan
5	0.31	Good	J	SI2	63.3	58	335
6	0.24	Very Good	J	VVS2	nan	57	336
7	0.24	Very Good	I	VVS1	62.3	57	336
8	0.26	Very Good	H	SI1	61.9	55	337
9	0.22	Fair	E	VS2	65.1	61	nan
10	0.23	Very Good	H	VS1	59.4	61	338
11	0.3	Good	J	SI1	64	55	339
12	0.23	Ideal	J	VS1	62.8	56	340
13	0.23	Ideal	J	VS1	nan	56	340
14	0.22	Premium	F	SI1	60.4	61	342
15	0.31	Ideal	J	SI2	62.2	54	344

- Import the appropriate Python libraries to create a Pandas DataFrame named `rubies` having the above columns. The columns and rows of the DataFrame should have appropriate names.

- Draw box plots for all numerical columns of the dataset in the same chart. Display the median of all numerical attributes in `rubies` for each type of `cut`.

- Display the per carat average price of all rubies grouped by the two attributes `clarity` and `color`.

- Normalize all quantitative features in range of [0,1].

- Draw word cloud for attribute `cut`.