SHIVAJI COLLEGE, UNIVERSITY OF DELHI
DEPARTMENT OF COMPUTER SCIENCE
INTERNAL TEST (Academic Year 2023-24)

Name of the Course : GE Semester: II
Name of the Paper : Data Analysis and Visualization using Python
Duration : Maximum Marks: 20 Date of Test : ARPRIL 22, 2024

SET-1

1. Consider a Dataframe. Write python commands to do the following (4)

One	Two	Three	Four	Five
1	14	34	NaN	NaN
34	21	NaN	12	NaN
NaN	23	NaN	2	NaN
34	21	32	33	NaN

   a. Drop columns with any null values
   b. Replace the null values with the mean of each column.
   c. Drop the null values where there are at least 2 null values in a row.
   d. Replace all null values by the last known valid observation

2. Consider the dataframe. Give commands to the following task (12)

EmployeeID	Department	Salary	Age	Designation
1001	English	1000	23	Professor
1002	English	1002	34	Associate
1003	English	1004	39	Assistant
1004	English	1000	43	Assistant
1003	Maths	1004	34	Assistant
1004	Maths	1005	43	Associate
1001	Maths	1006	53	Associate
1002	Maths	1002	43	Professor

- Determine the number of observations/records and the number of attributes in the dataframe.
- Display the names of the attributes, row indexes, and data types of each attribute in the dataframe.
- Display the first 5 and last 5 records of the dataframe.
- Retrieve the values of the second column for the third and fourth records.
- Display a summary of the data distribution for all attributes in the dataframe.
- Compute the pairwise correlation between all attributes in the dataframe.

3. Do the following using PANDAS Series: (4)
   a. Create a series with 5 elements. Display the series sorted on index and also sorted on values separately
   b. Create a series with N elements with some duplicate values. Find the minimum and maximum ranks assigned to the values using ‘first’ and ‘max’ methods
   c. Display the index value of the minimum and maximum element of a Series
   d. Make the first entry NaN of the series

Abha Vasal

(ABHA VASAL)

SHIVAJI COLLEGE, UNIVERSITY OF DELHI
DEPARTMENT OF COMPUTER SCIENCE
INTERNAL TEST (Academic Year 2023-24)

Name of the Course : GE Semester: II
Name of the Paper : Data Analysis and Visualization using Python
Duration : Maximum Marks: 20 Date of Test : ARPRIL 22, 2024

SET 2

Question : Consider the following dataframe df. Write suitable Python command(s) in Pandas library:

	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

1. Display the number of rows and columns present in the DataFrame df? (8)
2. 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.
3. 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.
4. 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?

Question 2. Consider the following Series: (5)

SR=pd.Series([8, -2, 6, 4, 3, 0, 6, -1], index=[0, 3, 4, 6, 7, 10, 12, 14]) find the output of the following:

a) SR.rank()
b) SR.rank(method='first')
c) SR.rank(ascending = False, method='min')
d) SR.reindex(range(6),method= 'ffill')
e) SR.replace({-2:np.nan,6: 0})

Question 3 **Consider the dataframe below**

	one	two	three	four
Delhi	0	1	2	3
Calcutta	4	5	6	7

(ABHA VASAL)

SHIVAJI COLLEGE, UNIVERSITY OF DELHI
DEPARTMENT OF COMPUTER SCIENCE
INTERNAL TEST (Academic Year 2023-24)

Name of the Course : GE Semester: II
Name of the Paper : Data Analysis and Visualization using Python
Duration : Maximum Marks: 20 Date of Test : ARPRIL 22, 2024

Pune	8	9	10	11
Aligarh	12	13	14	15

What is the output of following commands (5)

a. data[: 2]
b. data.loc['Calcutta', ['two', 'three']]
c. data.iloc[2, [3, 0, 1]]
d. data.iloc[: , : 3][data.three > 5]
e. data.loc[: 'Pune', 'two']

Question 4. Give examples to show how to fill missing values in a dataframe (2)

Abha Vasal

(ABHA VASAL)

SHIVAJI COLLEGE, UNIVERSITY OF DELHI
DEPARTMENT OF COMPUTER SCIENCE
INTERNAL TEST (Academic Year 2023-24)

Name of the Course : GE Semester: II
Name of the Paper : Data Analysis and Visualization using Python
Duration : Maximum Marks: 20 Date of Test : ARPRIL 22, 2024

SET 3

1. Consider the given data with two numerical attributes, *Hours_studied* and Marks_obtained loaded in a dataframe df. (10)

Hours_studied	3	4.5	9	10	1.5	2	8	1	11	6
Marks_obtained	45	50	90	95	33	20	89	0	80	72

   A. Give a Pandas command for computing the covariance among *Hours_studied* and *Marks_obtained*.
   B. What is meant by correlation between a pair of numerical attributes?
   C. Give a Pandas command to compute the correlation coefficient between Hours_studied and Marks_obtained.
   D. What is the possible range of values obtained in correlation and how is it interpreted?
   E. Add a column rest_hour which is equal to *Hours_studied +3*

2. 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. (3)
3. Consider 2 dataframes df1 and df2. What will be the output of following (4)

df1

	b	c	d
Ohio	**0.0**	**1.0**	**2.0**
Texas	**3.0**	**4.0**	**5.0**
Colorado	**6.0**	**7.0**	**8.0**

df2

	b	d	e
Utah	**0.0**	**1.0**	**2.0**
Ohio	**3.0**	**4.0**	**5.0**
Texas	**6.0**	**7.0**	**8.0**
Oregon	**9.0**	**10.0**	**11.0**

- df1 + df2
- df2.loc[1, 'b'] = np.nan
- df1.add(df2, fill_value=0)
- df1.reindex(columns=df2.columns, fill_value=0)

4. Differentiate between sort_index and sort_values command with examples (3)

(ABHA VASAL)