Vivekanand College, Kolhapur (Autonomous)

Department of Computer Science

Question Bank

Data Science

1. Consider the given table: **Product**. Write the python code for the following:

Item	Company	Rupees	USD
TV	LG	12000	700
TV	VIDEOCON	10000	650
TV	LG	15000	800
AC	SONY	14000	750

   i. To create the data frame for the above table.
   ii. To display the maximum price of LG TV.
   iii. To display the Sum of all products.
   iv. To display the median of the USD of Sony products.

2. Imagine you have grown to like Bollywood movies recently. Now you want to predict which of these actor’s movies you should watch when a new one is released. Here is a movie review dataset from the past that might help. It consists of three attributes: movie name, leading actor in the movie, and its IMDB rating. [Note: assume that a better rating means a more watchable movie.] Use Central Tendencies to come with conclusions.

Leading actor	Movie name	IMDB rating (out of 10)
Irfan Khan	Knock Out	6.0
Irfan Khan	New York	6.8
Irfan Khan	Life in a … metro	7.4
Anupam Kher	Striker	7.1
Anupam Kher	Dirty Politics	2.6
Anil Kapoor	Calcutta Mail	6.0
Anil Kapoor	Race	6.6

3. Consider following array,
   x = np.array([5,10,20,30,40,50,60,75,85,95]) and perform and shows the operations sum, minimum, maximum and product of elements of array x.

4. Consider following two arrays and performs arithmetic operations (addition, subtraction, multiplication and division) on it.
   x = np.array([[1, 2], [3, 4]])
   y = np.array([[5, 6], [7, 8]])

5. Consider following array and performs arithmetic operations (addition, subtraction, multiplication, division and exponential) on it with any scalar value.
   x = np.array([[1, 2], [3, 4]])

6. Consider any following array,
   x = np.array([5,10,20,30,40,50,60,75,85,95]) and perform and shows the operations mean, sum, minimum, maximum and product of elements of array x.

7. Create a 2-D array called arr1 using arange() having 20 rows and 4 columns with start value = 0,step size 0.5 having.
   Write commands for the following:
   a) Find the sum of all elements.
   b) Find the sum of all elements row wise.
   c) Find the sum of all elements column wise. Notes
   d) Find the max of all elements.
   e) Find the min of all elements in each row.
   f) Find the mean of all elements in each row.
   g) Find the standard deviation column wise.

8. Create the following NumPy arrays:
   a) A 1-D array called zeros having 10 elements and all the elements are set to zero.
   b) A 1-D array called vowels having the elements 'a', 'e', 'i', 'o' and 'u'.
   c) A 2-D array called ones having 2 rows and 5 columns and all the elements are set to 1 and dtype as int.
   d) Find the dimensions, shape, size, data type of the items and item size of arrays zeros, vowels and ones.
   e) Reshape the array ones to have all the 10 elements in a single row.

9. Use nested Python lists to create a 2-D array called a1 having 4 rows and 4 columns and store the following data:

```
2,   2,   2,   2

3,   3,   3,   3

4,   4,   4,   4

5,   5,   5,   5
```

A 2-D array called a2 using arange() having 4 rows and 4 columns with start value = 5,step size 5.

a) Divide all elements of array a1 by 3.

b) Add the arrays a1 and a2.

c) Subtract a1 from a2 and store the result in a new array.

d) Multiply a1 and a2 element wise.

e) Do the matrix multiplication of a1 and a2 and store the result in a new array a3.

f) Divide a1 by a2.

g) Find the cube of all elements of a1 and divide the resulting array by 2.

h) Find the square root of all elements of a2 and divide the resulting array by 2. The result should be rounded to two places of decimals.

10. The following table contains an imaginary dataset of auto insurance providers and their ratings as provided by the latest three customers. Now if you had to choose an auto insurance provider based on these ratings, which one would you opt for? Use central tendencies and explain with valid reasons.

#	**Insurance provider**	**Rating (out of 10)**
1	GEICO	4.7
2	GEICO	8.3
3	GEICO	9.2
4	Progressive	7.4
5	Progressive	6.7
6	Progressive	8.9
7	USAA	3.8
8	USAA	6.3
9	USAA	8.1

11. Create the following DataFrame Sales containing year wise sales figures for five sales persons in INR. Use the years as column labels, and sales person names as row labels.

	2014	**2015**	**2016**	**2017**
Madhu	100.5	12000	20000	50000
Kusum	150.8	18000	50000	60000
Kinshuk	200.9	22000	70000	70000
Ankit	30000	30000	100000	80000
Shruti	40000	45000	125000	90000

a) Display the row labels of Sales.
b) Display the column labels of Sales.
c) Display the data types of each column of Sales.
d) Display the dimensions, shape, size and values of Sales.
e) Display the last two rows of Sales.
f) Display the first two columns of Sales.

12. Create a horizontal bar graph of following data. Add suitable labels.

City	**Population**
Delhi	23456123
Mumbai	20083104
Bangalore	18456123
Hyderabad	13411093

13. Write the python statement for the following question on the basis of given dataset:

	Name	Degree	Score
0	Aparna	MBA	90
1	Pankaj	BCA	NaN
2	Ram	M.Tech	80
3	Ramesh	MBA	98
4	Naveen	NaN	97
5	Krrishnav	BCA	78
6	Bhawna	MBA	89

a) To create the above DataFrame.
b) To print the Degree and maximum marks in each stream.
c) To fill the NaN with 76.
d) To set the index to Name.
e) To display the name and degree wise average marks of each student.
f) To count the number of students in MBA.
g) To print the mode marks BCA.

14. Download auto-mpg dataset from the UCI repository/Kaggle.
Following are the exercises to analyse the data.
1) Load auto-mpg.data into a DataFrame autodf.
2) Give description of the generated DataFrame autodf.
3) Display the first 10 rows of the DataFrame autodf.
4) Find the attributes which have missing values. Handle the missing values using following two ways:
i. Replace the missing values by a average value.
ii. Remove the rows having missing values from the original dataset
5) Print the details of the car which gave the maximum mileage.
6) Find the average displacement of the car given the number of cylinders.
7) What is the average number of cylinders in a car?
8) Determine the no. of cars with weight greater than the average weight.

15. Download Titanic dataset from the UCI repository/Kaggle.
Following are the exercises to analyse the data.
1) Load Titanic.csv into a DataFrame JackRose.
2) Give information of the generated DataFrame JackRose.
3) Display the last 10 rows of the DataFrame JackRose.
4) Describe the dataframe JackRose.
5) Find the attributes which have missing values and display the count.
6) Show the percentage of missing values for each attribute.
7) Display the male and female count of passengers.
8) Display class wise survived count of passengers.

16. Download Titanic dataset from the UCI repository/Kaggle.

    Following are the exercises to analyse the data.

    1) Load Titanic.csv into a DataFrame JackRose.
    2) Give information of the generated DataFrame JackRose.
    3) Display shape of the DataFrame JackRose.
    4) Count the class wise passengers and show it with pie chart.
    5) Show distribution of the attribute AGE using histogram.
    6) Count the passengers who have paid more than 500 dollars.
    7) Display the count of survived passengers who have paid more than 400 dollars.
    8) Show the scatter plot for age of passenger with fare paid by them.

17. Consider the average heights and weights of persons aged 8 to 16 stored in the following two lists:

    height = [121.9,124.5,129.5,134.6,139.7,147.3, 152.4, 157.5,162.6]
    weight= [19.7,21.3,23.5,25.9,28.5,32.1,35.7,39.6, 43.2]

    Let us plot a line chart where:

    i. x axis will represent weight
    ii. y axis will represent height
    iii. x axis label should be "Weight in kg"
    iv. y axis label should be "Height in cm"
    v. colour of the line should be green
    vi. use * as marker
    vii. Marker size as10
    viii. The title of the chart should be "Average weight with respect to average height".
    ix. Line style should be dashed
    x. Linewidth should be 2.

18. Smile NGO has participated in a three week cultural mela. Using Pandas, they have stored the sales (in Rs) made day wise for every week in a CSV file named "MelaSales.csv", as shown in Table

Week 1	**Week 2**	**Week 3**
5000	4000	4000
5900	3000	5800
6500	5000	3500
3500	5500	2500
4000	3000	3000
5300	4300	5300
7900	5900	6000

Depict the sales for the three weeks using a Line chart. It should have the following:

i. Chart title as "Mela Sales Report".
ii. axis label as Days.
iii. axis label as "Sales in Rs".
iv. Line colours are red for week 1, blue for week 2 and brown
for week 3.

19. Create Mela.csv and draw the Bar plot with column Day on X-axis and Sales in Rs for each week on Y-axis.

Week 1	Week 2	Week 3	Day
5000	4000	4000	Monday
5900	3000	5800	Tuesday
6500	5000	3500	Wednesday
3500	5500	2500	Thursday
4000	3000	3000	Friday
5300	4300	5300	Saturday
7900	5900	6000	Sunday

20. Prayatna sells designer bags and wallets. During the sales season, he gave discounts ranging from 10% to 50% over a period of 5 weeks. He recorded his sales for each type of discount in an array. Draw a scatter plot to show a relationship between the discount offered and sales made. (Discount on X-axis and Sales on Y-axis)

21. Let us consider the dataset of Table showing the forest cover of north eastern states that contains geographical area and corresponding forest cover in sq km along with the names of the corresponding states. Draw pie charts for showing Geographical area and Forest cover area respectively.

State	GeoArea	ForestCover
Arunachal Pradesh	83743	67353
Assam	78438	27692
Manipur	22327	17280
Meghalaya	22429	17321
Mizoram	21081	19240
Nagaland	16579	13464
Tripura	10486	8073

22. Consider the following dataframe, and answer the questions given below:

```
import pandas as pd
df = pd.DataFrame({“Quarter1":[2000, 4000, 5000, 4400, 10000],
 "Quarter2":[5800, 2500, 5400, 3000, 2900],
 "Quarter3":[20000, 16000, 7000, 3600, 8200],
 "Quarter4":[1400, 3700, 1700, 2000, 6000]})
```

i) Write the code to find mean value from above dataframedf over the index and column axis. (Skip NaN value)

ii) Use sum() function to find the sum of all the values over the index axis.

23. What is exploratory analysis? Explain.

24. Define term Population and explain its types.

25. Create a horizontal bar graph of following data. Add suitable labels.

City	**Population**
Delhi	23456123
Mumbai	20083104
Bangalore	18456123
Hyderabad	13411093

26. Define Machine learning? Briefly explain the types of learning.
27. Write a note on Series and DataFrame in Pandas.
28. What is the role of Drew Conway's Venn diagram in data science?
29. Write a note on Open Data Principles.
30. What is Data Science? Where do we see Data Science?
31. How does data science relate to other fields?
32. Explain skills required for Data Science?
33. Write a note on tools for Data Science.
34. Explain various types of data?
35. Write a note on data collection.
36. Explain the data pre-processing in detail.
37. What is the role of Drew Conway's Venn diagram in data science?
38. Explain the data science process in detail.
39. Define term Population and explain its types.
40. Define term Sample and explain about its properties.
41. Define term Sampling Unit, Sampling Frame and Sampling Error.
42. Write a note on Open Data Principles.
43. Write a note on EDA.
44. Write a note on EDA techniques and tools.
45. Write a note on structured with two examples.
46. Write a note on unstructured data with two examples.
47. Write and explain basic properties of an ndarray class.
48. Write any five functions to create an array using NumPy in python.
49. Write any five array indexing and slicing expression in NumPy.

50. Consider following two arrays and performs arithmetic operations (addition, subtraction, multiplication and division) on it.
    x = np.array([[1, 2], [3, 4]])
    y = np.array([[5, 6], [7, 8]])

51. Consider following array and performs arithmetic operations (addition, subtraction, multiplication, division and exponential) on it with any scalar value.
    x = np.array([[1, 2], [3, 4]])

52. Consider any following array,
    x = np.array([5,10,20,30,40,50,60,75,85,95]) and perform and shows the operations mean, sum, minimum, maximum and product of elements of array x.

53. Write a note on Series and DataFrame in Pandas.

54. Explain following DataFrame attributes with examples.
Index, Columns, Values, Shape, Size.

55. Explain following Series methods with examples.
head, count, tail

56. What are missing values? What are the strategies to handle them?