| RegisterNumber | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Set-C

**SRM Institute of Science and TechnologyCollegeofEngineeringandTechnology**
**SchoolofComputing**

SRMNagar,Kattankulathur–603203,ChengalpattuDistrict,TamilNadu

**AcademicYear:2024-25(Even)**

Test:FT1 Date:25-02-2025
CourseCode&Title:21CSS303T-Data Science Duration:50 Minutes
Year& Sem: IIIYear /VISem Max.Marks:25

CourseArticulationMatrix:

| Course Outcome | PO1 | PO2 | PO3 | PO4 | PO5 | PO6 | PO7 | PO8 | PO9 | PO10 | PO11 | PO12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CO1 | - | - | - | - | 1 | - | - | - | - | - | - | - |
| CO2 | - | - | - | - | 1 | - | - | - | - | - | - | - |

**Note:** CO1 - To understand the relationship between data
CO2 - Identify the different data structures to represent data

**Part– A**

(5x2= 10 Marks)

Answer ALL the questions

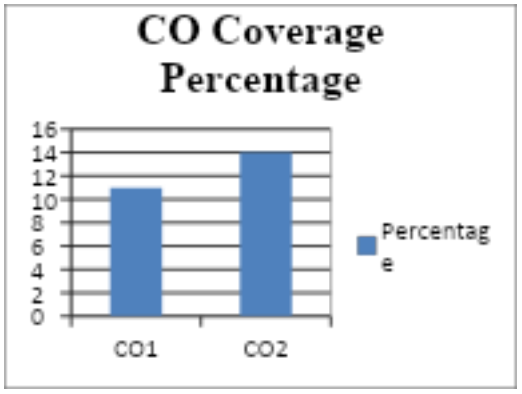
| Q.No | Question | Marks | BL | CO | PO | PI.Code |
|---|---|---|---|---|---|---|
| 1 | What is the goal of the "exploratory data analysis" phase? | 2 | 1 | 1 | 5 | 5.6.1 |
| 2 | Write the syntax to create a 1D NumPy array from a Python list. | 2 | 1 | 1 | 5 | 5.6.1 |
| 3 | Why are NumPy arrays more efficient than Python lists for numerical operations? | 2 | 2 | 1 | 5 | 5.4.1 |
| 4 | Compare a Python list and a Pandas Series? | 2 | 2 | 2 | 5 | 5.4.1 |
| 5 | How would you display the first five rows of a DataFrame? | 2 | 2 | 2 | 5 | 5.4.1 |

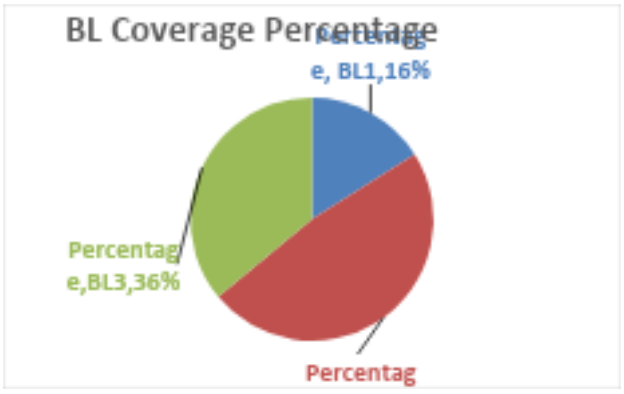
**Part– B**

(3x5= 15 Marks)

| Q.No | Question | Marks | BL | CO | PO | PI.Code |
|---|---|---|---|---|---|---|
| 1 | Explain the different facets of data in Data Science with suitable examples. | 5 | 2 | 1 | 5 | 5.4.2 |
| 2 | Given the following dataset stored in sales_data.csv:<br>Product \| Category \| Sales<br>A \| Electronics \| 1000<br>B \| Clothing \| 500<br>C \| Electronics \| 1200<br>D \| Clothing \| 700<br>E \| Grocery \| 300<br>Write a Python program to:<br>Read the CSV file into a DataFrame<br>Find the total sales per category<br>Find the average sales per category | 5 | 3 | 2 | 5 | 5.5.1 |
| 3 | Explain different types of data acquisition techniques used in Data Science. | 5 | 2 | 2 | 5 | 5.4.2 |

**Course Outcome (CO)andBloom's level (BL)Coverage in Questions**




## Key:

**1. What is the goal of the "exploratory data analysis" phase?**

Exploratory Data Analysis (EDA) is an important first step in data science.Its goal is to gain insights by looking at and visualizing data to understand its main features, find patterns, spotting anomalies, validating assumptions and discover how different parts of the data are connected before applying any machine learning models or statistical techniques. **2M**

**2. Write the syntax to create a 1D NumPy array from a Python list.**

```
import numpy as np
// Creating a 1D NumPy array from a Python list
my_list = [1, 2, 3, 4, 5]     1M
np_array = np.array(my_list)  1M
print(np_array)
```

**3. Why are NumPy arrays more efficient than Python lists for numerical operations?**

NumPy is faster and more memory-efficient than Python lists because of contiguous memory storage, vectorized operations(operations are applied to all elements in an array without the need for explicit loops in Python), broadcasting, and optimized C-based backend (uses BLAS (Basic Linear Algebra Subprograms) and LAPACK(Linear Algebra PACKage), which are highly optimized C libraries) computations. **Any two explanations each 1M**

**4. Compare a Python list and a Pandas Series?**

| Feature | Lists | Pandas Series |
|---|---|---|
| Missing Values | Must handle manually | Built-in support for NaN |
| Performance | Slower | Faster |
| Memory Usage | Higher | Lower |
| Indexing | Uses integer-based indexing | Supports custom indexing |
| **Any two difference each 1M** | | |

**5. How would you display the first five rows of a DataFrame?**

first five rows of a Pandas DataFrame can be displayed using the .head() method. 2M

**Part B**

**1. Explain the different facets of data in Data Science with suitable examples.**

Very large amount of data will generate in big data and data science. These data is various types and main categories of data are as follows:
a) Structured
b) Natural language
c) Graph-based
d) Streaming
e) Unstructured
f) Machine-generated
g) Audio, video and images each 1Mark with appropriate explanation (any five)

2.

Given the following dataset stored in sales_data.csv:

| Product | Category | Sales |
|---|---|---|

| | | |
|---|---|---|
| A | Electronics | 1000 |
| B | Clothing | 500 |
| C | Electronics | 1200 |
| D | Clothing | 700 |
| E | Grocery | 300 |

Write a Python program to:
Read the CSV file into a DataFrame
Find the total sales per category
Find the average sales per category
**Ans:**

```
import pandas as pd
df = pd.read_csv("sales_data.csv")
# Create a DataFrame
df = pd.DataFrame(data)
# Find the total sales per category
total_sales = df.groupby("Category")["Sales"].sum()
# Find the average sales per category
average_sales = df.groupby("Category")["Sales"].mean()
# Display results
print("Total Sales per Category:")
print(total_sales)
print("\nAverage Sales per Category:")
print(average_sales)
```

**1 mark -reading csv file**
**2 mark -total sales per category**
**2 mark - average sales**

## 3. Explain different types of data acquisition techniques used in Data Science.

Ans: Data Science primarily involve methods to collect raw data from various sources, including sensors, databases, APIs, and manual inputs

Methods of different data collection includes primary data ans secondary data.
Primary data:

- Direct Personal Investigation:
- Indirect Oral Investigation:
- Information from Local Sources or Correspondents
- Information through Questionnaires and Schedules
- Mailing Method
- Enumerator's Method

**Any 3 methods with explanation 3 x 1 =3M**

Secondary data

- Published Sources (Government Publications,Semi-Government Publications,Publications of Trade Associations,Journals and Papers,International Publications,Publications of Research Institutions)
- Unpublished Sources (These organizations usually collect data for their self-use and are not published anywhere.)
- Web Scraping

**Any two with explanation 2 x1 = 2M**