INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal-500043, Hyderabad

B.Tech V SEMESTER END EXAMINATIONS (REGULAR) - DECEMBER 2022
Regulation:UG20
DATA WRANGLING WITH PYTHON

Time: 3 Hours   (CSE-DATA SCIENCE)   **Max Marks: 70**

Answer ALL questions in Module I and II
Answer ONE out of two questions in Modules III, IV and V
All Questions Carry Equal Marks
All parts of the question must be answered in one place only

MODULE – I

1. (a) Identify the importance and core activities of data wrangling. Create a list and show different ways of accessing a list with an example. [BL: Understand| CO: 1|Marks: 7]

   (b) For the dataset given in Table 1, using python, add a few rows to the data frame and then show how to remove duplicate rows from the same data frame in different methods

   [BL: Apply| CO: 1|Marks: 7]

Table 1

	Name	Surname	Salary	Locality	Gender
0	ABC	BC	15.0	LOC1	0.0
1	DEF	TH	14.0	NAN	0.1
2	XYZ	CV	13.0	LOC3	0.0

MODULE – II

2. (a) Discuss the methodology through a flowchart for inserting a data object into a non-relational databse using MongoClient. Also provide python code. [BL: Understand| CO: 2|Marks: 7]

   (b) Write a python program to write the data into a CSV file given in Table 2.

Table 2

Year	Category	Film	Winner
2015	Visual Effects	Mad Max Fury Road	Richard Stammers
2014	Best Picture	The Wolf of Wall Street Martin	Scorsese
2013	Music	Skyfall	Adele Adkins
2012	Cinematography	X-Men: Days of Future Past	Doug Mitchell

[BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) Discuss the possible causes of missing values, bad data, and duplicate values in raw data. Also review the possible solutionsl [BL: Understand| CO: 3|Marks: 7]

    (b) Create a dataset, perform i) Filtering of outliers ii) Handling missing data using python, and write down the expected outcome. [BL: Apply| CO: 3|Marks: 7]

4. (a) How to determine outliers in the given data? Summarize the Fuzzy matching process of data cleaning. [BL: Understand| CO: 4|Marks: 7]

    (b) Create a dataset and perform i) Removing of duplicates ii) Fixing of structural errors.
    Illustarte i) Valid data, ii) Accurate data, iii) consistent data, iv) unique data using python and write down the output regarding the same. [BL: Apply| CO: 4|Marks: 7]

MODULE – IV

5. (a) What do you mean by time series data? How can we plot it? Explain it with example to plot trend over time. [BL: Understand| CO: 5|Marks: 7]

    (b) Create two tables and use table functions to merge/join tables with the help of python.
    [BL: Apply| CO: 5|Marks: 7]

6. (a) Interpret the use of scatter-plot in data visualization. Can we draw trendline in scatter-plot? Explain it with an example. [BL: Understand| CO: 5|Marks: 7]

    (b) Using the dataset given in Table 2, visualize the dataset using i) Pie chart ii) Bar chart iii) Histogram iv) Waterfall chart [BL: Apply| CO: 5|Marks: 7]

Table 3

Month	January	February	March	April	May	June
Spending expenditure	3000	5000	3000	6000	2000	5000
Month	July	August	September	October	November	December
Spending expenditure	2500	4000	6000	8000	8250	10000

MODULE – V

7. (a) How do you scrape and analyze a web page? Write a detailed user-centric list of website analysis methods. [BL: Understand| CO: 6|Marks: 7]

    (b) Develop a python program to read a web page with suitable tags and parameters.
    [BL: Apply| CO: 6|Marks: 7]

8. (a) What is browser-based parsing? Demonstrate the browser-based parsing of a web page.
    [BL: Understand| CO: 6|Marks: 7]

    (b) Consider a Html file. Write python program to scrap the page and extract values associated with tags and properties. [BL: Apply| CO: 6|Marks: 7]

− ∘ ∘ ◯ ∘ ∘ −