Hall Ticket No □□□□□□□□□□ Question Paper Code: ACDC05

INSTITUTE OF AERONAUTICAL ENGINEERING

(Autonomous)
Dundigal-500043, Hyderabad

B.Tech V SEMESTER END EXAMINATIONS (REGULAR/ SUPPLEMENTARY) - FEBRUARY 2024

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) Outline the basics of Python programming, including data structures, control flow and functions, essential for performing data wrangling operations effectively.
[BL: Understand| CO: 1|Marks: 7]

   (b) Analyze how effective data wrangling practices contribute to enhancing data quality, improving analysis outcomes, and informing data-driven decision-making processes?
[BL: Apply| CO: 1|Marks: 7]

MODULE – II

2. (a) Classify different approaches to PDF parsing, including text extraction, metadata extraction, and structural analysis. [BL: Understand| CO: 2|Marks: 7]

   (b) Discuss best practices for installing and managing Python packages using tools such as pip and conda, and explore strategies for resolving dependencies and ensuring compatibility across different environments. [BL: Apply| CO: 2|Marks: 7]

MODULE – III

3. (a) Why is it essential to clean data before analysis or modeling? List the potential consequences of working with dirty or uncleaned data. [BL: Understand| CO: 3|Marks: 7]

   (b) Describe the fundamental principles and techniques of data cleanup, and how do they contribute to enhancing the quality of datasets? [BL: Understand| CO: 3|Marks: 7]

4. (a) How can Python be leveraged to perform data cleanup tasks efficiently and effectively?
[BL: Understand| CO: 4|Marks: 7]

   (b) Explain the strategies for documenting data cleanup processes and collaborating with stakeholders to validate and refine cleanup strategies. [BL: Understand| CO: 4|Marks: 7]

MODULE – IV

5. (a) What are the key considerations for presenting and visualizing data effectively using Python based tools and libraries? [BL: Understand| CO: 5|Marks: 7]

   (b) Explore the diverse range of visualization options available, including charts, time-related data visualizations, maps, and interactive visualizations. [BL: Apply| CO: 5|Marks: 7].

6. (a) Identify correlations, outliers, and groupings within datasets using Python? What insights can be derived from these analyses? [BL: Understand| CO: 5|Marks: 7]

   (b) Summarize the functionality provided by pandas DataFrame methods, including filtering, sorting, grouping, and aggregation. Elucidate how these functions enable comprehensive data exploration and analysis? [BL: Apply| CO: 5|Marks: 7]

MODULE – V

7. (a) Demonstrate the architecture and functionality of scrapy, including its capabilities for managing crawling behavior, handling concurrency, and scaling to large-scale web scraping tasks. [BL: Understand| CO: 6|Marks: 7]

   (b) Elucidate the best practices for configuring and optimizing scrapy spiders to maximize scraping efficiency and minimize detection and blocking risks. [BL: Apply| CO: 6|Marks: 7]

8. (a) How can Python frameworks such as scrapy be leveraged to build robust web spiders for crawling and scraping whole websites? [BL: Understand| CO: 6|Marks: 7]

   (b) Infer strategies for screen reading and dynamic content extraction in web scraping workflows. [BL: Understand| CO: 6|Marks: 7]

– ○ ○ ◯ ○ ○ –