DATA ANALYSIS AND DATA SCIENCE USING PYTHON

TASK - 1

Task 1: Data Analysis Project Using Python

Objective:
Analyze a dataset of student exam scores and answer specific questions, presenting findings using Python libraries and techniques for data analysis.

Project Steps

1. Dataset Selection

- **Dataset**: Download the **Student Performance Dataset** from a provided source.
  - File: `student-mat.csv`.
  - Contains columns like:
    - `G1, G2, G3` (grades for three terms).
    - `study time` (hours spent studying weekly).
    - `sex` (gender: Male/Female).

2. Tasks to Perform

a. Data Loading

- Load the dataset using **pandas**.
- Display the first few rows using `.head()`.

b. Data Exploration

- Check for missing values using `.isnull().sum()`.
- Display column data types using `.dtypes`.
- Understand the dataset's size using `.shape`.

c. Data Cleaning

- Handle missing values (e.g., replace them with the median or remove rows).
- Remove duplicate entries using `.drop_duplicates()`.

d. Data Analysis Questions

1. **What is the average score in math (`G3`)?**
2. **How many students scored above 15 in their final grade (`G3`)?**
3. **Is there a correlation between study time (`study time`) and the final grade (`G3`)?**
4. **Which gender has a higher average final grade (`G3`)?**

Main Flow Services and Technologies Pvt. Ltd.
Contact Us. +91 9389641586, +91 97736 99074
Email-Add. contact.mainflow@gmail.com
www.mainflow.in

e. Data Visualization

1. Plot a **histogram of final grades (`G3`)**.
2. Create a **scatter plot** between study time (`study time`) and final grade (`G3`).
3. Create a **bar chart** comparing the average scores of male and female students.

Restrictions

1. **Data Loading and Cleaning**
   - Use only basic pandas operations for loading, cleaning, and manipulating data.
   - **Reason**: To teach fundamental data handling and exploration without over-reliance on pre-built functions or external tools.
2. **Analysis and Calculations**
   - Perform all calculations (e.g., averages, correlations) using pandas and NumPy, without third-party statistical libraries like `scipy`.
   - **Reason**: To focus on understanding mathematical concepts rather than using pre-built statistical packages.
3. **Visualization**
   - Use only **matplotlib** or **seaborn** for plotting. Avoid high-level tools like Plotly for simplicity.
   - **Reason**: To ensure students learn basic plotting techniques before advancing to interactive visualizations.
4. **Code Format**
   - The code should be written in a **single Jupyter Notebook** with clear cell divisions.
   - **Reason**: Encourages organized and modular programming, simulating professional data science practices.
5. **Documentation**
   - Provide explanations in Markdown cells for each step of the analysis and visualizations.
   - **Reason**: Teaches students how to communicate findings effectively and add context to their code.
6. **Deadline**
   - Must be submitted **within 7 days**.
   - **Reason**: Reinforces discipline and time management, simulating real-world project deadlines.

Deliverables

1. **Python Code**:
   - Well-documented Jupyter Notebook with structured code and step-by-step analysis.

Main Flow Services and Technologies Pvt. Ltd.
Contact Us. +91 9389641586, +91 97736 99074
Email-Add. contact.mainflow@gmail.com
www.mainflow.in

2. **Analysis Summary**:
    - Markdown cells explaining:
        - The purpose of each step.
        - Key findings from the data analysis and visualizations.

Evaluation Criteria

1. **Code Quality**:
    - Is the code organized, well-commented, and efficient?
2. **Data Insights**:
    - Are the analysis questions answered correctly with supporting calculations?
3. **Visualizations**:
    - Are the visualizations clear, relevant, and well-labeled?
4. **Report Quality**:
    - Does the Markdown documentation effectively explain the process and findings?

Learning Outcomes

- Master data exploration, cleaning, and manipulation using pandas.
- Understand basic statistical calculations and correlation concepts.
- Develop skills to create meaningful visualizations.
- Learn to document findings and insights clearly and effectively.

Deadline Compliance

- **Restriction**: **Submit the project within 7 days** from the start date.
- **Reason**: Meeting deadlines is crucial in the real-world software development environment. This restriction helps students practice **time management** and **task prioritization**. In professional settings, tight deadlines are often the norm, and learning to meet them without compromising quality is an essential skill.
- **Learning Outcome**: Students will learn to manage their time effectively, complete projects under pressure, and **deliver results on time**, which are all important skills in the workplace.