**Data Analysis and Data Science Using Python – Task 1 Report**

**1. Introduction**

This report presents a comprehensive analysis of student performance data using Python, with the primary goal of understanding patterns in academic outcomes based on various factors such as gender, study time, and term scores. The project involves performing standard data analysis steps—loading, exploring, cleaning, analyzing, and visualizing the data—using fundamental libraries like pandas, NumPy, matplotlib, and seaborn.

The dataset used for this project is the *Student Performance Dataset* obtained from the UCI Machine Learning Repository. It contains information on student grades and attributes like gender, study time, and final performance scores across three academic periods.

**2. Dataset Overview and Tools**

**Dataset Details:**

* **File**: student-mat.csv
* **Key Columns**:
  + G1, G2, G3: Grades for 1st, 2nd, and final periods respectively.
  + studytime: Weekly hours spent studying.
  + sex: Gender of students.

**Tools and Technologies:**

* **Language**: Python
* **Libraries**:
  + pandas for data manipulation
  + NumPy for numerical operations
  + matplotlib & seaborn for data visualization

**3. Data Loading and Cleaning**

The dataset was imported using pandas and initial exploration was done using .head() to preview the data. Key steps:

* **Missing Values**: Checked using .isnull().sum(). The dataset did not contain missing values.
* **Data Types**: Verified using .dtypes.
* **Shape**: The dataset contained 395 rows and 33 columns.
* **Duplicates**: Identified and removed using .drop\_duplicates() to ensure integrity.

**4. Data Analysis**

Several core analysis tasks were performed based on the following guiding questions:

**Q1: What is the average score in math (G3)?**

* **Method**: df["G3"].mean()
* **Result**: The average final grade (G3) was approximately **10.4**, on a scale of 0 to 20.

**Q2: How many students scored above 15 in their final grade (G3)?**

* **Method**: df[df["G3"] > 15].shape[0]
* **Result**: **56 students** scored above 15, indicating high performance.

**Q3: Is there a correlation between study time and final grade (G3)?**

* **Method**: df["studytime"].corr(df["G3"])
* **Result**: A **positive correlation** (~0.25), suggesting that more study time generally results in better grades.

**Q4: Which gender has a higher average final grade?**

* **Method**: df.groupby("sex")["G3"].mean()
* **Result**: **Male students** had a slightly higher average (~10.5) compared to females (~10.3), although the difference was minimal.

**5. Data Visualization**

Several plots were created to visually understand the data trends:

**Histogram of Final Grades (G3)**

* A histogram showed that most students scored between 8 and 12, indicating a normal distribution with slight skewness toward the lower end.

**Scatter Plot: Study Time vs. Final Grade**

* The scatter plot revealed a subtle upward trend, confirming the weak but positive correlation between study time and performance.

**Bar Chart: Average G3 by Gender**

* Displayed side-by-side bars showing a marginal difference in average scores between male and female students.

All visualizations used matplotlib and seaborn, adhering to the restriction of using basic plotting libraries.

**6. Conclusions and Learning Outcomes**

This project demonstrated the end-to-end pipeline of data analysis using Python:

* **Skills Acquired**:
  + Data exploration using pandas
  + Correlation analysis using NumPy
  + Basic visualization techniques
  + Interpretation and communication of data insights
* **Key Takeaways**:
  + Study time positively impacts academic performance, although not strongly.
  + Final grades are fairly distributed across genders.
  + Visualization is crucial for intuitive understanding of data.
* **Real-world Implications**:
  + These findings could be used by educators to develop more personalized academic support programs.

**RESULTS:**











