Group Assignment Submission Instructions

Please follow the instructions below to complete and submit our group assignment:

1. **Group Members:**
   Ensure that your group consists of exactly five members. Include the full names of all members on the first page of your submission.

2. **Assignment Tasks:**

   - **Write and execute the Python code required for the assigned questions.**

   - **Capture clear screenshots of your code and the corresponding outputs/results for each question.**

   - **Provide concise but complete written answers or explanations alongside your code where applicable.**

3. **Documentation:**

   - **Compile all screenshots, code snippets, and answers into a single Microsoft Word document.**

   - **Organize the document logically with clear headings for each question or task.**

   - **Include a cover page with the group member names, assignment title, and date.**

4. **File Conversion:**

   - **Save or export the Word document as a PDF file to ensure formatting is preserved.**

5. **Submission:**

- **Email the PDF file to joseptuyish@gmail.com.**
- **Use the subject line: Group Assignment Submission – [Your Group Name].**
- **Double-check that the PDF is attached before sending.**

6. **Deadline:**
   - **Submit your assignment by 16/02/2026 at 20:25. Late submissions may not be accepted or may incur penalties.**

Student_ID	Age	Score	Hours_Studied
S1	20	85	10
S2	21		15
S3		78	7
S4	19	90	
S5	22	88	12
S6	twenty	92	9
S7	23	NaN	14
S8	24	87	11
S9		80	8

Tasks

1. Create a pandas DataFrame with the above data.
2. View the dataset and summarize its contents.
3. Identify columns with missing or invalid data.
4. Convert the Age and Score columns to numeric types, treating invalid entries (e.g., 'twenty', 'NaN') as missing values.
5. Treat all missing values in numeric columns by replacing them with the mean of their respective columns.
6. Plot a scatter plot showing the relationship between Hours_Studied (x-axis) and Score (y-axis).

    - ➢ Include a meaningful title and axis labels.

YearsExperience	**MonthlySalary**	**Team**	**EducationLevel**	**RemoteWorkStatus**
3	450	Development	Bachelor's	Yes
5	5200	Marketing	Master's	No
NaN	4800	Development	None	Yes
2	NaN	Sales	Bachelor's	Maybe
4	5100	Marketing	Bachelor's	Yes
NaN	5300	Sales	Master's	No

a) Describe the structure of the dataset, including the types of variables it contains and any potential data quality issues you observe, such as missing or inconsistent values. Describe how understanding these aspects is important before performing further analysis.

b) Write Python code to calculate and print the median monthly salary, ensuring that missing or invalid salary data are handled properly to produce an accurate statistic

c) Since the dataset lacks a performance score column, generate one by assigning random integer scores between 50 and 100 to each employee. Then, create a scatter plot showing Years of Experience versus Performance Score, coloring points by Remote Work Status. Use descriptive statistics to summarize the relationship between Years of Experience and Performance Score

dex	**YearsExperience**	**Salary**
0	1.1	39343
1	2.0	46205
2	3.2	*NaN*
3	4.5	60000
4	*NaN*	65200
5	6.8	72500
6	7.5	*NaN*
7	8.3	83000
8	9.0	88000
9	10.5	95000

a) Train a simple linear regression model on the training data by initializing the regressor and fitting it with the training feature and target values. Simple linear regression is appropriate here because it models the relationship between one independent variable and a continuous dependent variable, allowing you to predict the target based on the given feature

b) Use the trained simple linear regression model to predict values on the test set and then compute the mean squared error (MSE) or another appropriate regression metric. This metric measures the average squared difference between the predicted and actual values, providing insight into the model's prediction accuracy.

c) Extract and print the coefficient (slope) and intercept of the simple linear regression model. Interpret the sign of the coefficient: a positive coefficient means that as the independent variable increases, the target variable tends to increase, whereas a negative coefficient indicates an inverse relationship. The magnitude of the coefficient shows the strength of this effect.