■ Monthly Exam: Impact of Digital Learning on Student Performance

■ Problem Statement

In recent years, digital learning and online education have rapidly increased — especially after the COVID-19 pandemic.
Education institutions now want to analyze how different factors — such as gender, parental education, lunch type, test preparation, and study mode (online or offline) — influence student academic performance.

You are hired as a Data Analyst / Data Scientist to explore this data, visualize insights, and build a predictive or clustering model to understand student behavior and performance trends in the digital learning era.

■ Dataset

Use the dataset:
■ Students Performance in Exams | Kaggle
https://www.kaggle.com/datasets/spscientist/students-performance-in-exams

You can assume there is an additional column:
study_mode → with values "online" or "offline" (you may simulate or add it manually for analysis).

■ Exam Tasks

1■■ Data Loading & Preprocessing
- Load the dataset and show its first 5 rows.
- Check for missing values and handle them appropriately.
- Add a column for average_score = (Math + Reading + Writing) / 3.
- (Optional) Categorize performance as "High", "Average", "Low" based on average score.
- Convert categorical columns (like gender, study_mode) into numeric form if needed.

2■■ Exploratory Data Analysis (EDA)
Perform data exploration and answer with plots + observations:
- Compare the average performance of students in online vs offline study modes.
- How do parental education and test preparation affect performance?
- Which gender performs better in Math, Reading, and Writing?
- Visualize the distribution of scores for each subject (histogram / boxplot).
- Check correlations between numeric variables using a heatmap.
- Identify any outliers or extreme patterns in the dataset.
- Find which features are most associated with high performance.

3■■ Machine Learning Task
Choose one approach:

■ A. Supervised Learning
- Predict whether a student’s performance is High, Average, or Low.
- Use algorithms like Logistic Regression, Decision Tree, or Random Forest.
- Evaluate using accuracy, confusion matrix, and classification report.

- Identify which features most strongly impact student success.

■ B. Unsupervised Learning

- Apply K-Means Clustering to group students based on performance and demographics.
- Find the optimal number of clusters (Elbow Method).
- Visualize clusters using 2D plots.
- Interpret what each cluster represents (e.g., high-performing online learners, low-performing offline learners, etc.).

4■■ Final Insights & Recommendations

- Summarize major findings from your analysis.
- Describe how digital learning mode affects student performance.
- Suggest data-driven strategies for improving learning outcomes.
- Mention limitations of the dataset and possible improvements for future analysis.