# **📝 ML Engineering Assignment (90 Minutes)**

## **Problem Statement**

A retail company wants to predict whether a customer will make a **purchase** after visiting their website. You are provided with a small dataset (customer\_visits\_training\_data.csv) containing basic customer and session information.

The dataset is intentionally **messy** to simulate real-world conditions — it contains missing values, inconsistent data formats, and outliers.

Your task is to **clean, preprocess, build models, and recommend the best approach**.

## **Dataset Description**

The dataset has ~250 rows with the following columns:

* customer\_id — Unique identifier (string, may contain duplicates)
* age — Customer age (int, but some missing or unrealistic values like -5, 120)
* country — Country (string, but some rows have typos like Inda, US, united states)
* pages\_visited — Number of pages visited in the session (int, some negative values)
* time\_spent — Time spent on site in minutes (float, some missing values)
* purchase — Target variable (0 = No Purchase, 1 = Purchase)

## **Your Tasks (90 Minutes)**

### **1. Data Exploration & Cleaning (25 min)**

* Identify and handle discrepancies.
* Summarize how you handled discrepancies and why.

### **2. Preprocessing (20 min)**

* Encode categorical features (country).
* Normalize/scale numeric features if needed.
* Split dataset into train/test (80/20).

### **3. Modeling (30 min)**

* Train two models (e.g., Logistic Regression and Decision Tree / Random Forest).
* Evaluate using: Accuracy, Precision, Recall, F1-score.
* Compare performance
* Use the customer\_visits\_predict.csv data and predict the purchase using the selected model

### **4. Communication (15 min)**

* Recommend the better model with justification.
* Explain how data quality issues impacted the results.
* Suggest two improvements for data collection/cleaning in the future.

## **Constraints**

* No external internet searches, pre-written solutions, or AI-generated code.
* Allowed libraries: pandas, numpy, scikit-learn, matplotlib (optional for simple plots).
* Focus on **clarity, correctness, and reasoning**.

## **Deliverables**

1. A **Python script or Jupyter Notebook** with your solution.
2. A **short written summary** (5–8 sentences) with:  
   * Cleaning decisions
   * Model recommendation
   * Future improvements