# Data Cleaning Challenge Assignment

Objective: Clean and prepare a messy dataset for use in an AI project

You have received a new dataset – but this time, it's in CSV format instead of JSON. The structure is slightly different from what we used in class, and the dataset contains various issues. Your task is to apply your knowledge to:

- Understand and analyze the data
- Identify and fix problems in the dataset
- Make decisions on how to handle them
- Prepare the dataset so it can be used in an AI model

## Instructions

1. Load the CSV file (`messy_data.csv`) into a pandas DataFrame.

2. Explore the data:
   - What values are missing?
   - Are there any unrealistic values? (e.g., age = 150, height = -99)
   - Are there format or type issues that need fixing?

3. Clean the data step-by-step:
   - Handle missing values (you decide how)
   - Identify and process outliers (you define the thresholds)
   - Normalize relevant numerical columns (e.g., height, weight, income)
   - Convert categorical values to numerical format using one-hot encoding
   - Ensure all height values are valid (e.g., no negatives)

4. Convert the `purchases` column:
   - This column is a string – convert it into a list of numbers
   - Replace missing values inside the list (e.g., empty strings) with 0

5. Save your cleaned DataFrame as a new CSV file (`cleaned_data.csv`).

## Optional Challenge

If you finish early, try adding:
- A function that automatically detects outliers in a column
- A plot showing the distribution of age, income, or number of purchases