**Name:Garige.Kalpana**

**Roll Number:2503B05117**

**Lab 17 – AI for Data Processing: Data Cleaning and Preprocessing Scripts**

**Task 1 – Social Media Data Cleaning**

**Task: Clean raw social media posts dataset.**

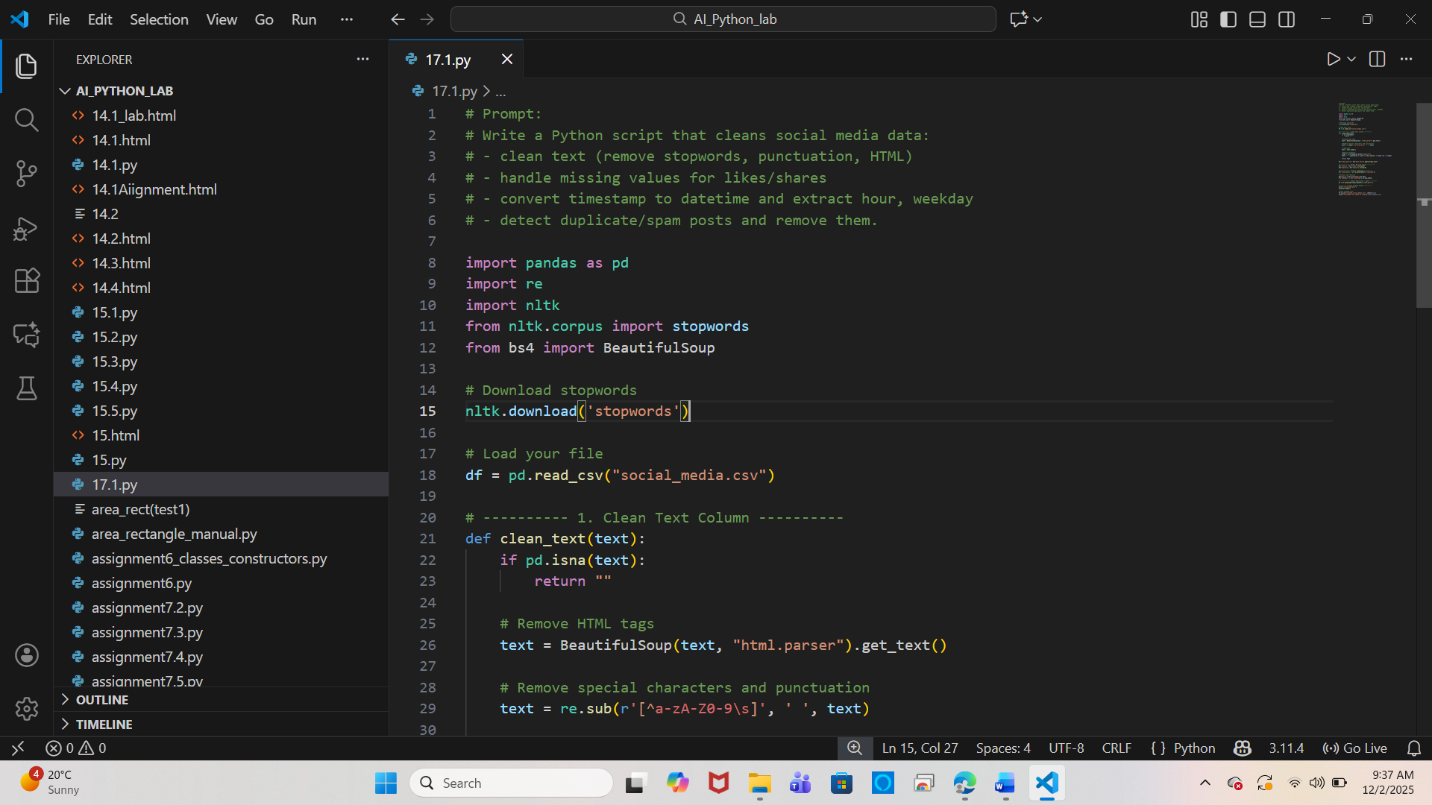
**Instructions:**

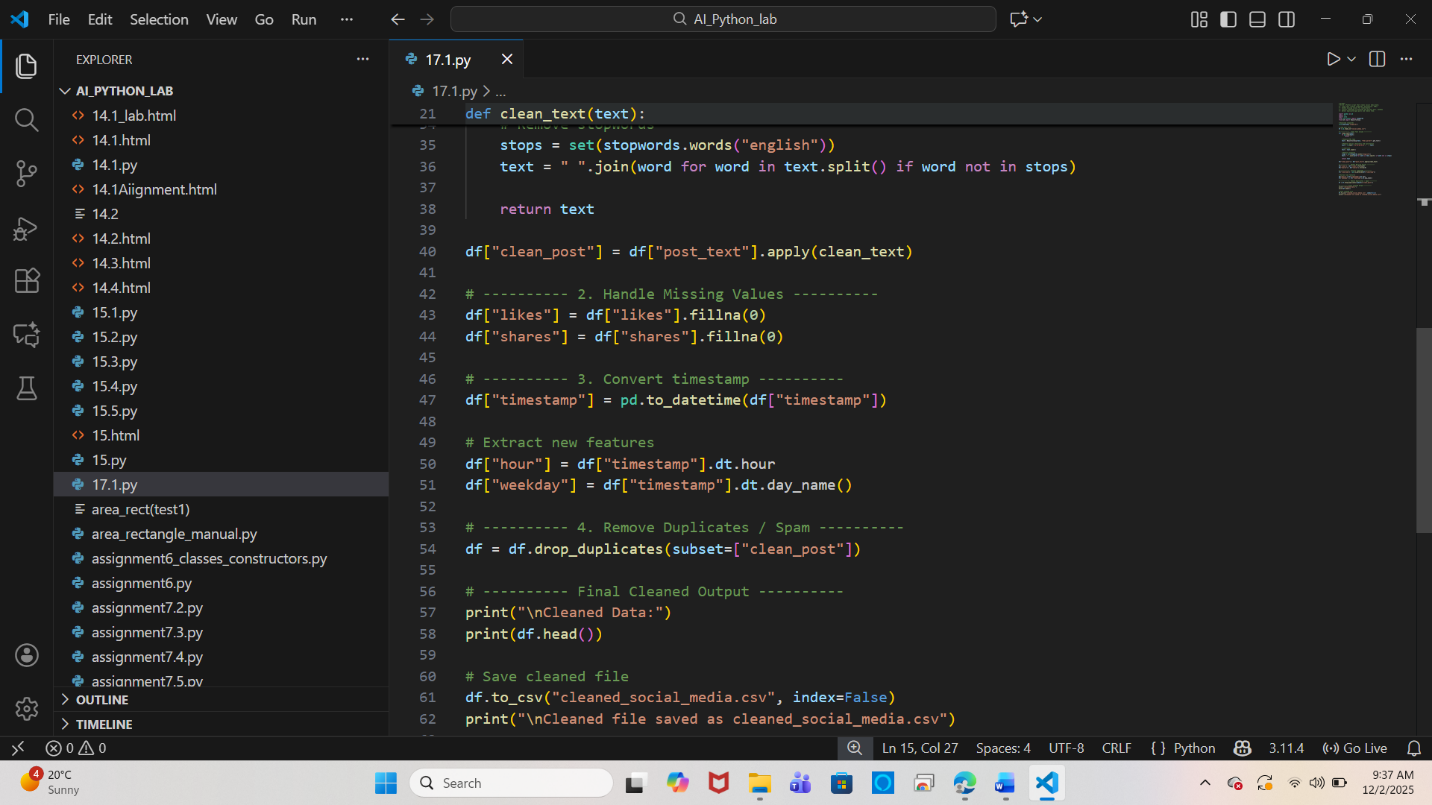
**- Remove stopwords, punctuation, and special symbols from post text.**

**- Handle missing values in likes and shares columns.**

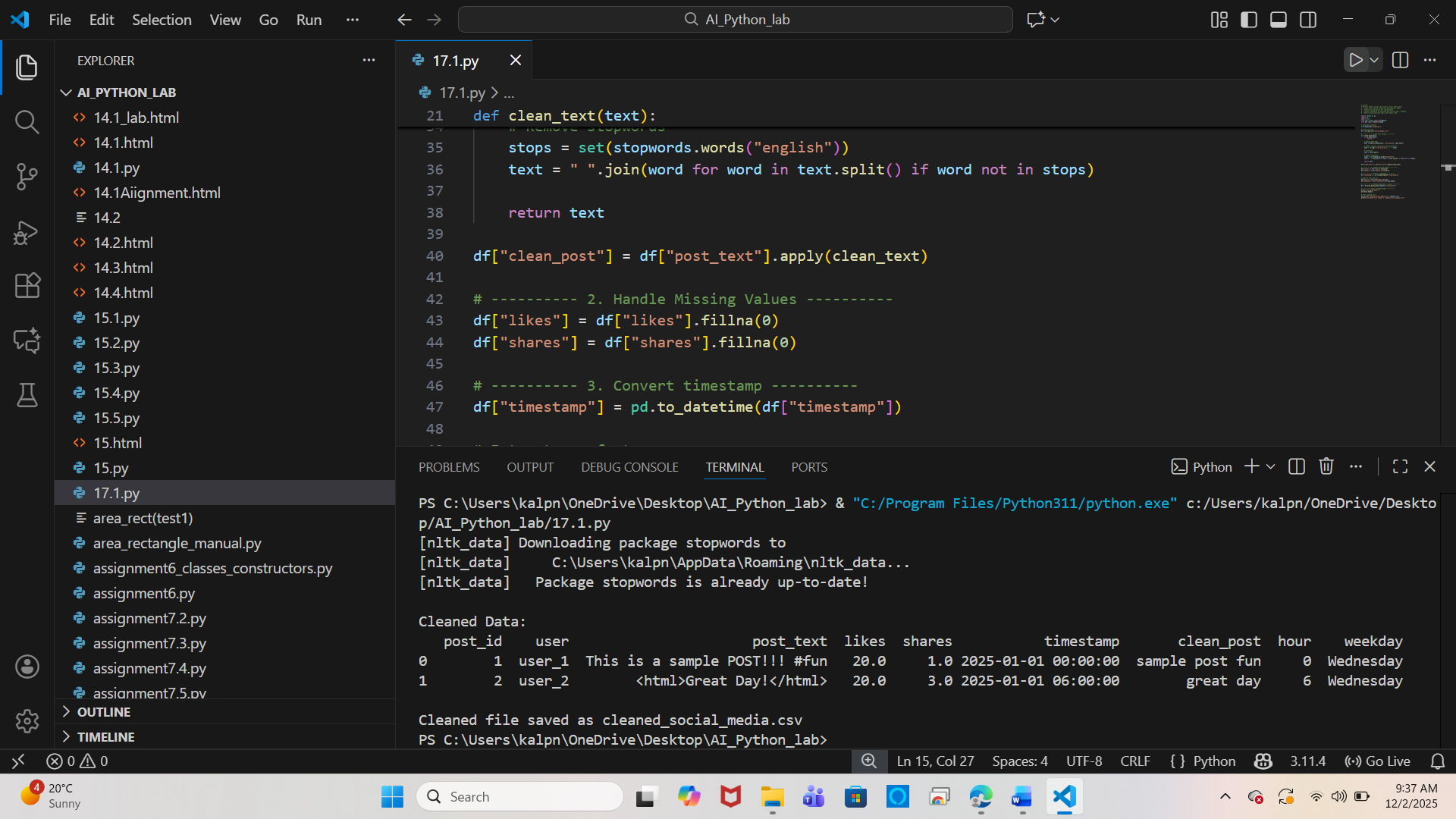
**- Convert timestamp to datetime and extract features (hour, weekday).**

**- Detect and remove spam/duplicate posts.**

****

****

**Output:**

****

**Task 2 – Financial Data Preprocessing**

**Task: Preprocess a stock market dataset.**

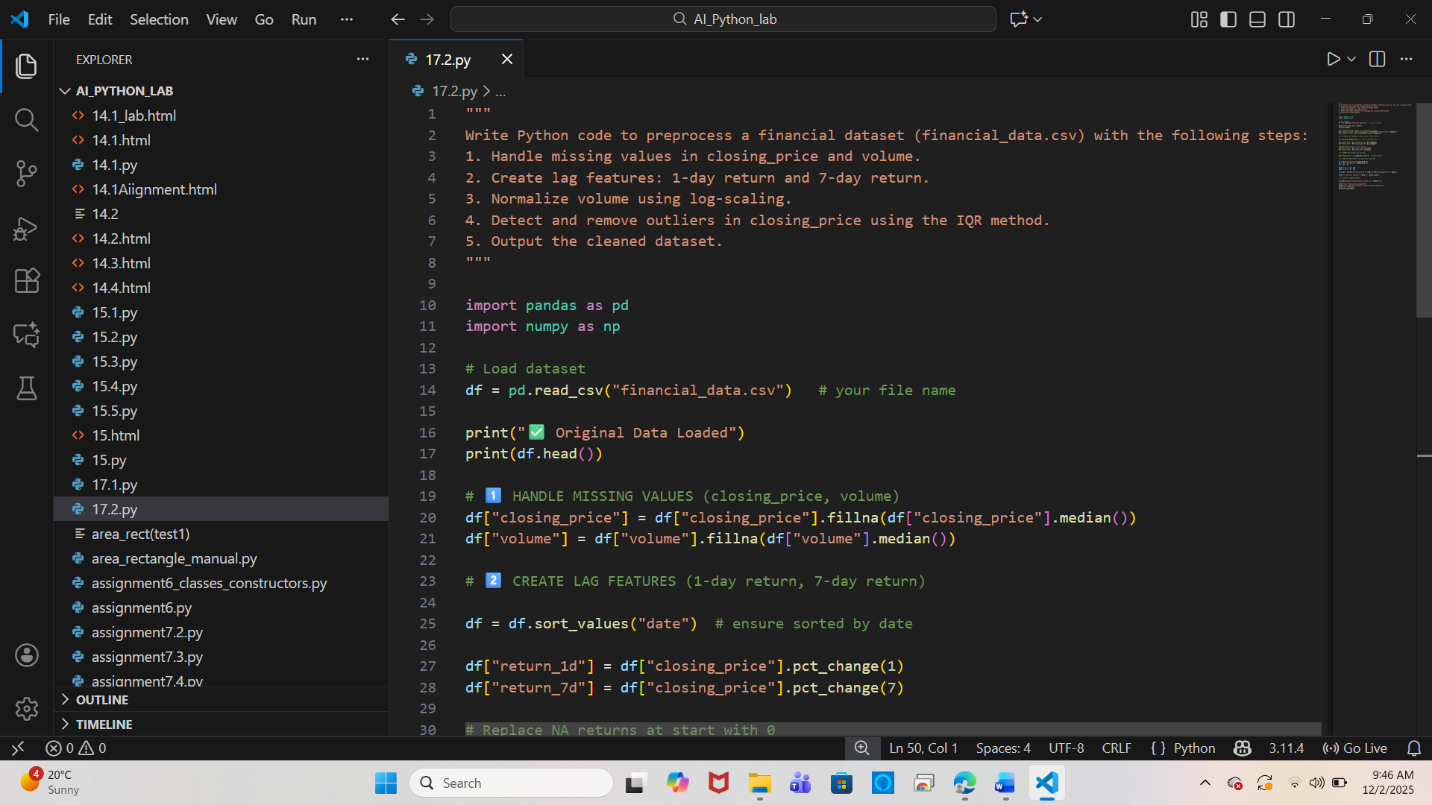
**Instructions:**

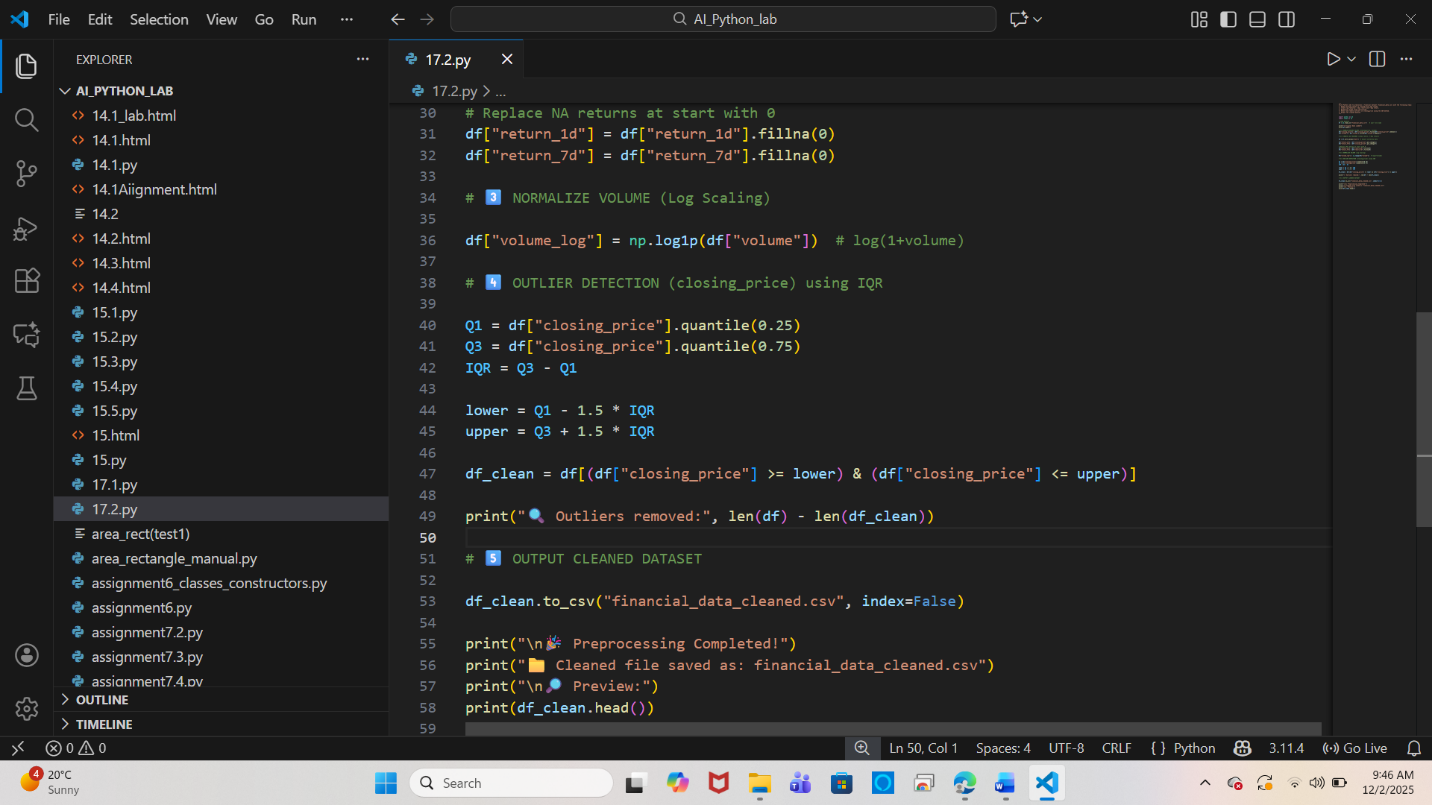
**- Handle missing values in closing\_price and volume.**

**- Create lag features (1-day, 7-day returns).**

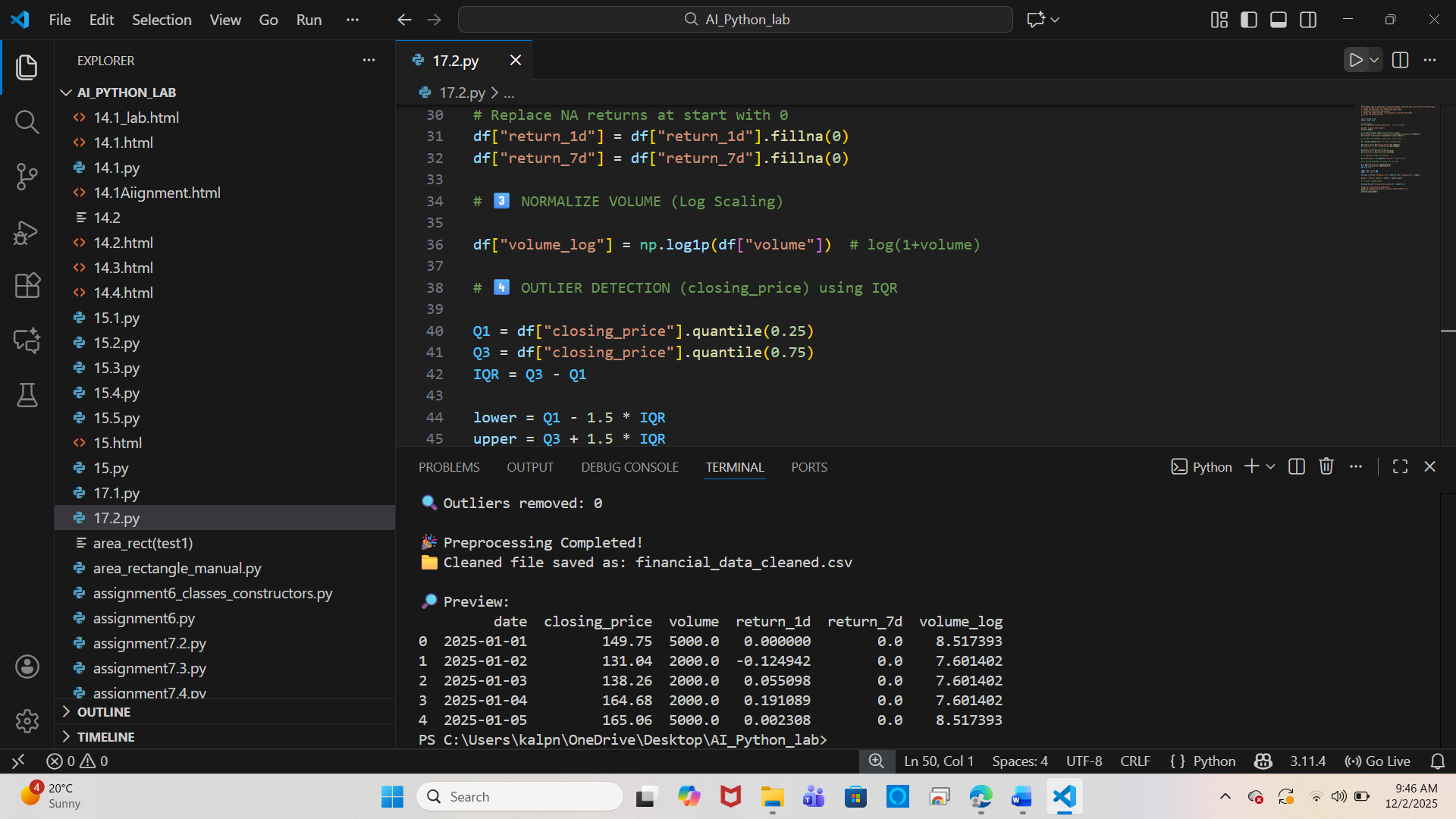
**- Normalize volume column using log-scaling.**

**- Detect outliers in closing\_price using IQR method.**

****

****

**Output:**

****

**Task 3 – IoT Sensor Data Preparation**

**Task: Clean and preprocess IoT temperature and humidity logs.**

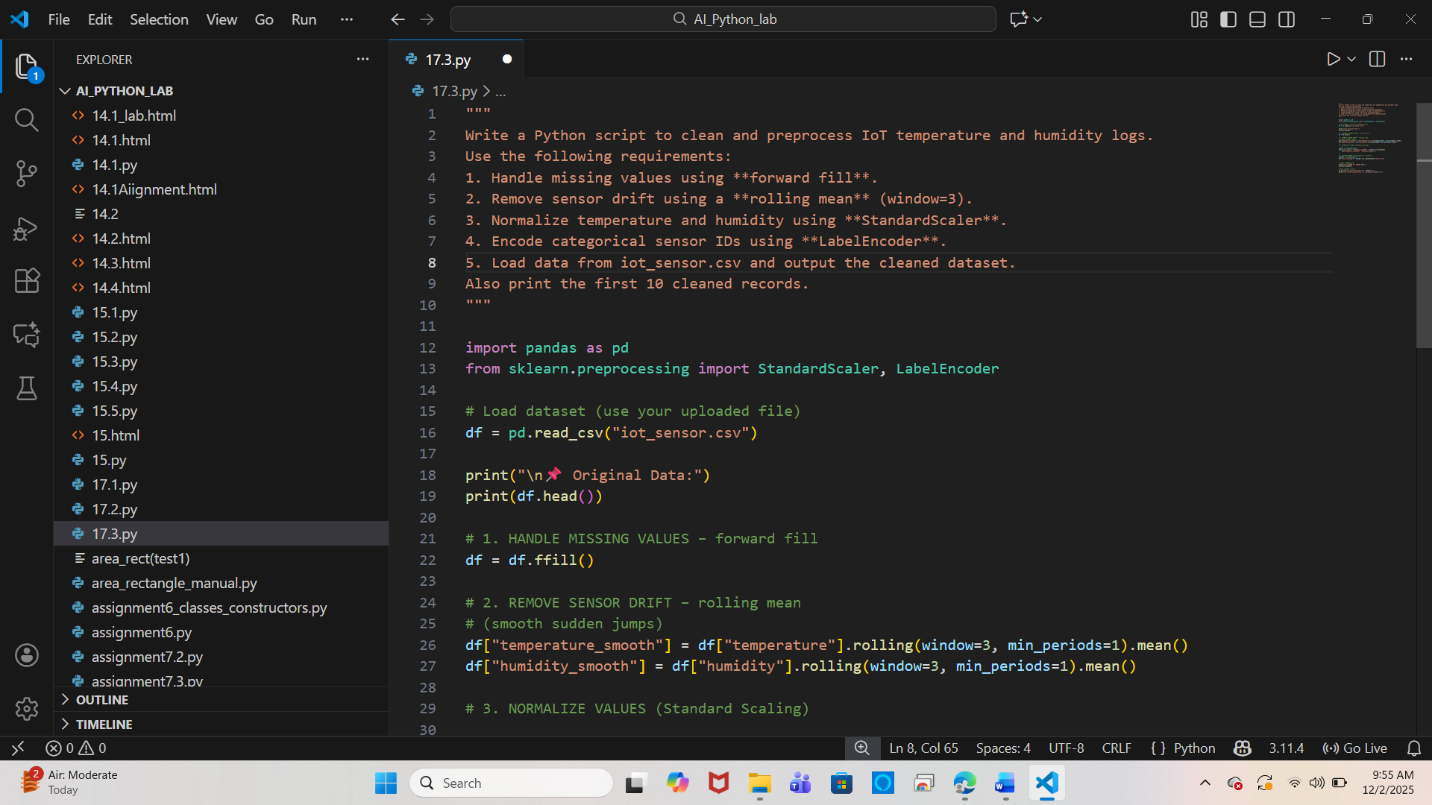
**Instructions:**

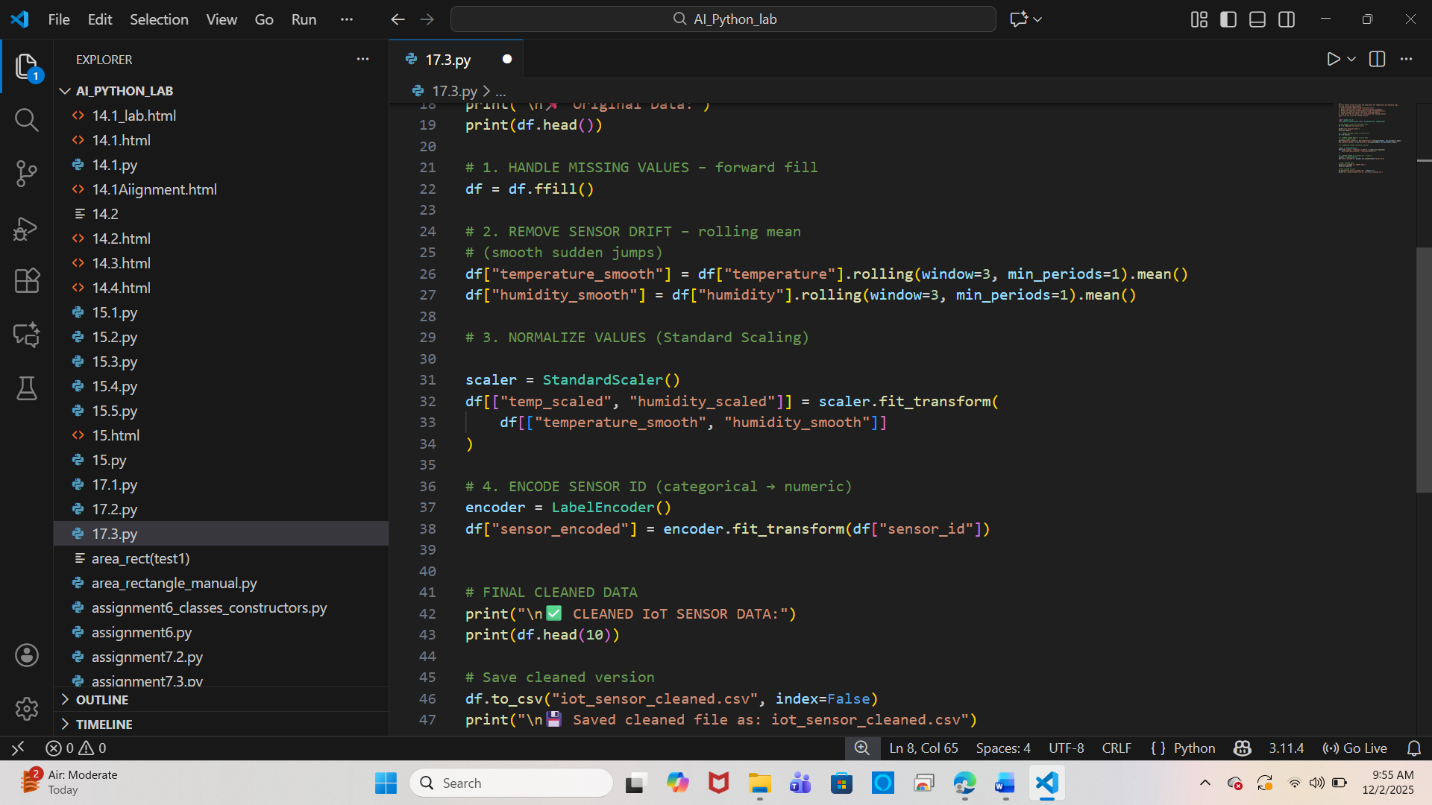
**- Handle missing values using forward fill.**

**- Remove sensor drift (apply rolling mean).**

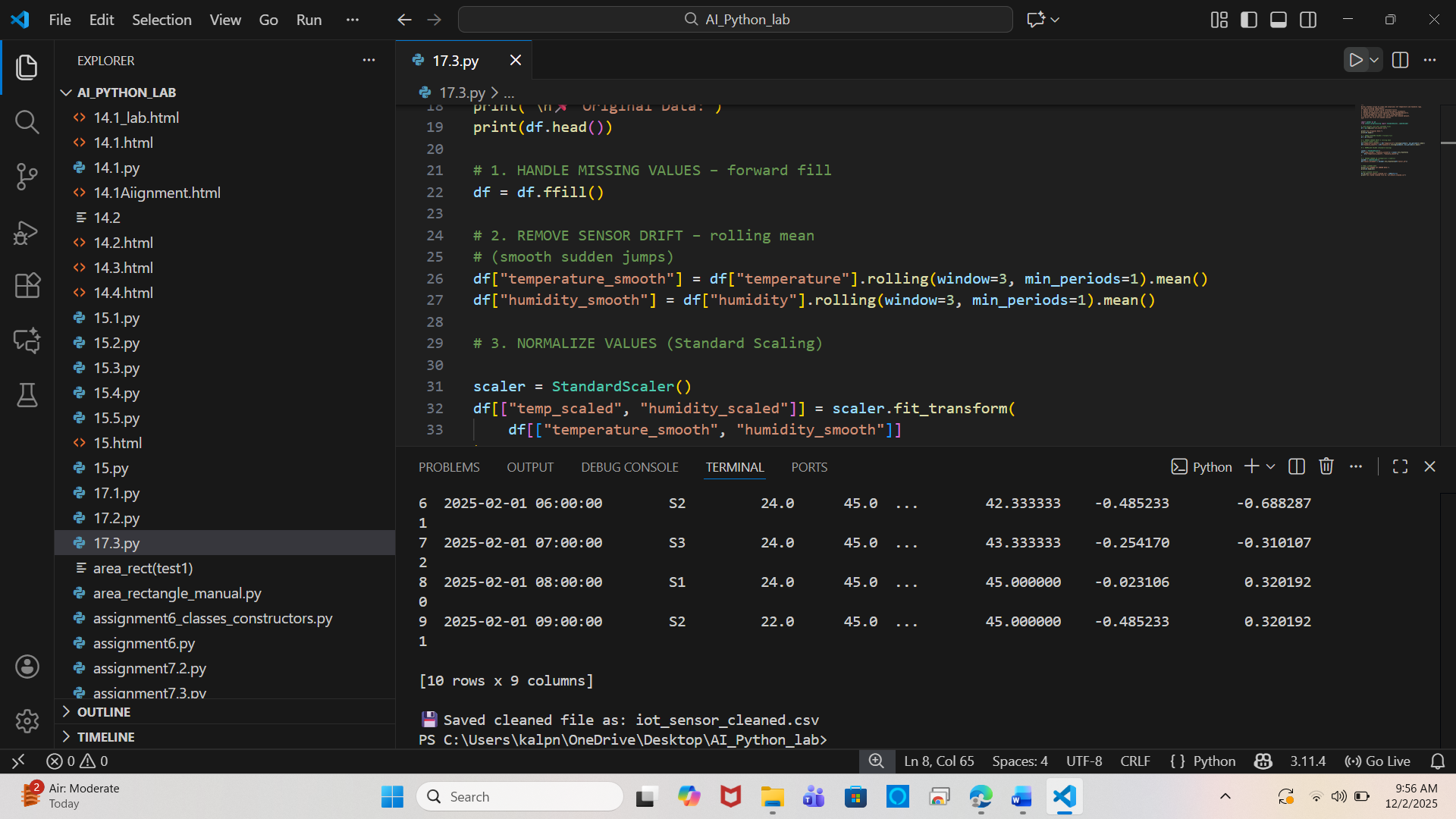
**- Normalize readings using standard scaling.**

**- Encode categorical sensor IDs.**

****

****

**Output:**

****

**Task 4 – Real-Time Application: Movie Reviews Data Cleaning**

**Task: A streaming platform wants to analyze customer reviews.**

**Instructions:**

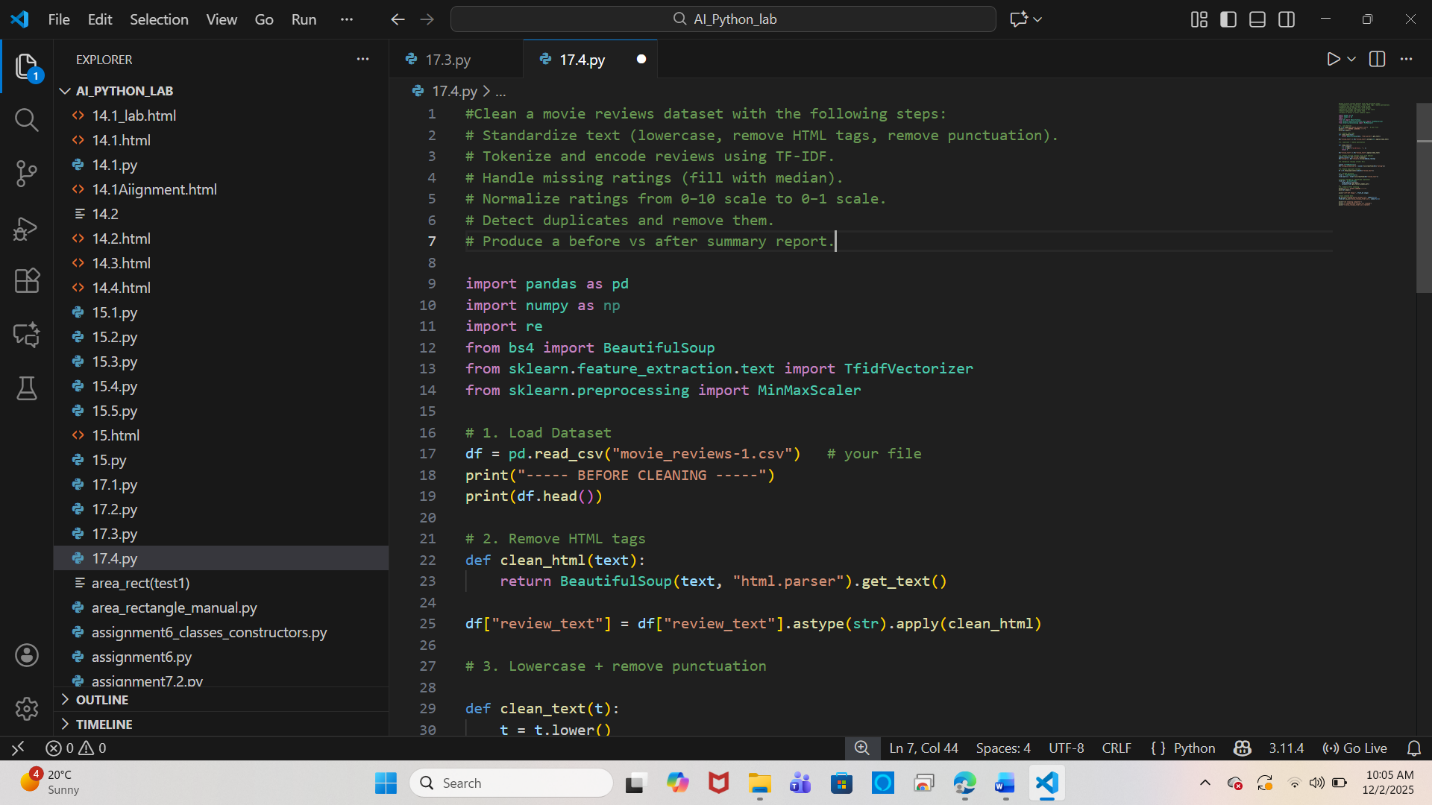
**- Standardize text (lowercase, remove HTML tags).**

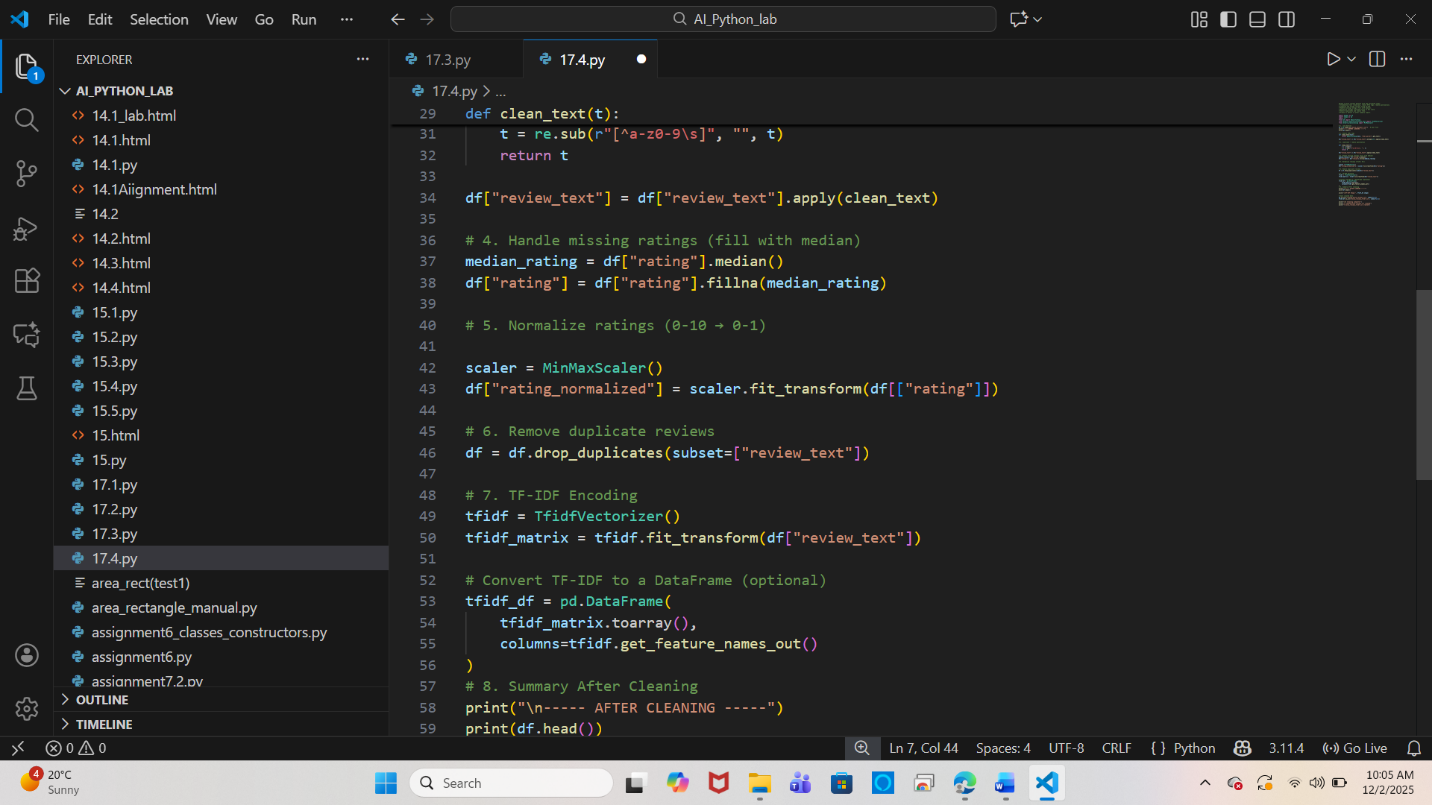
**- Tokenize and encode reviews using AI-assisted methods (TF-IDF or embeddings).**

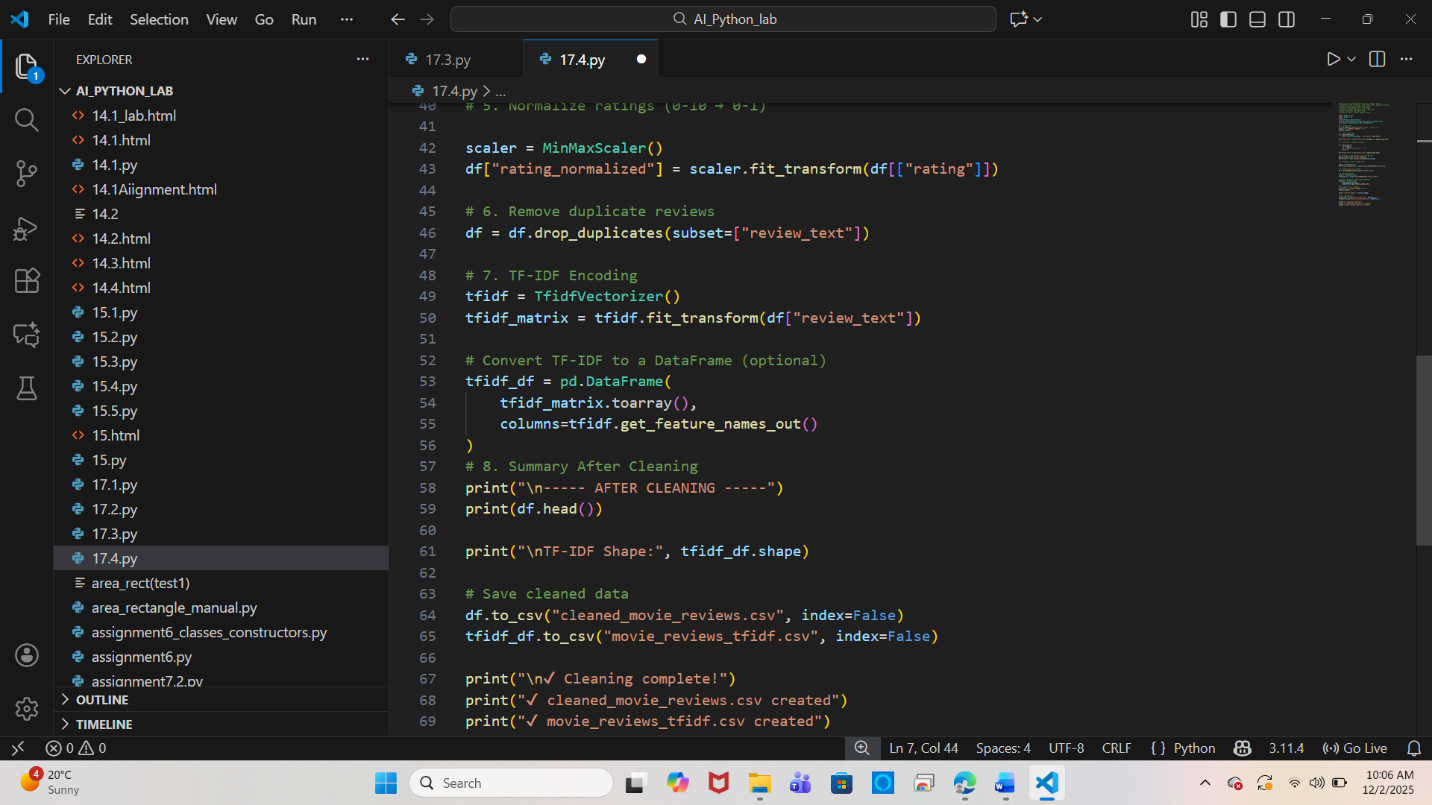
**- Handle missing ratings (fill with median).**

**- Normalize ratings (0–10 → 0–1 scale).**

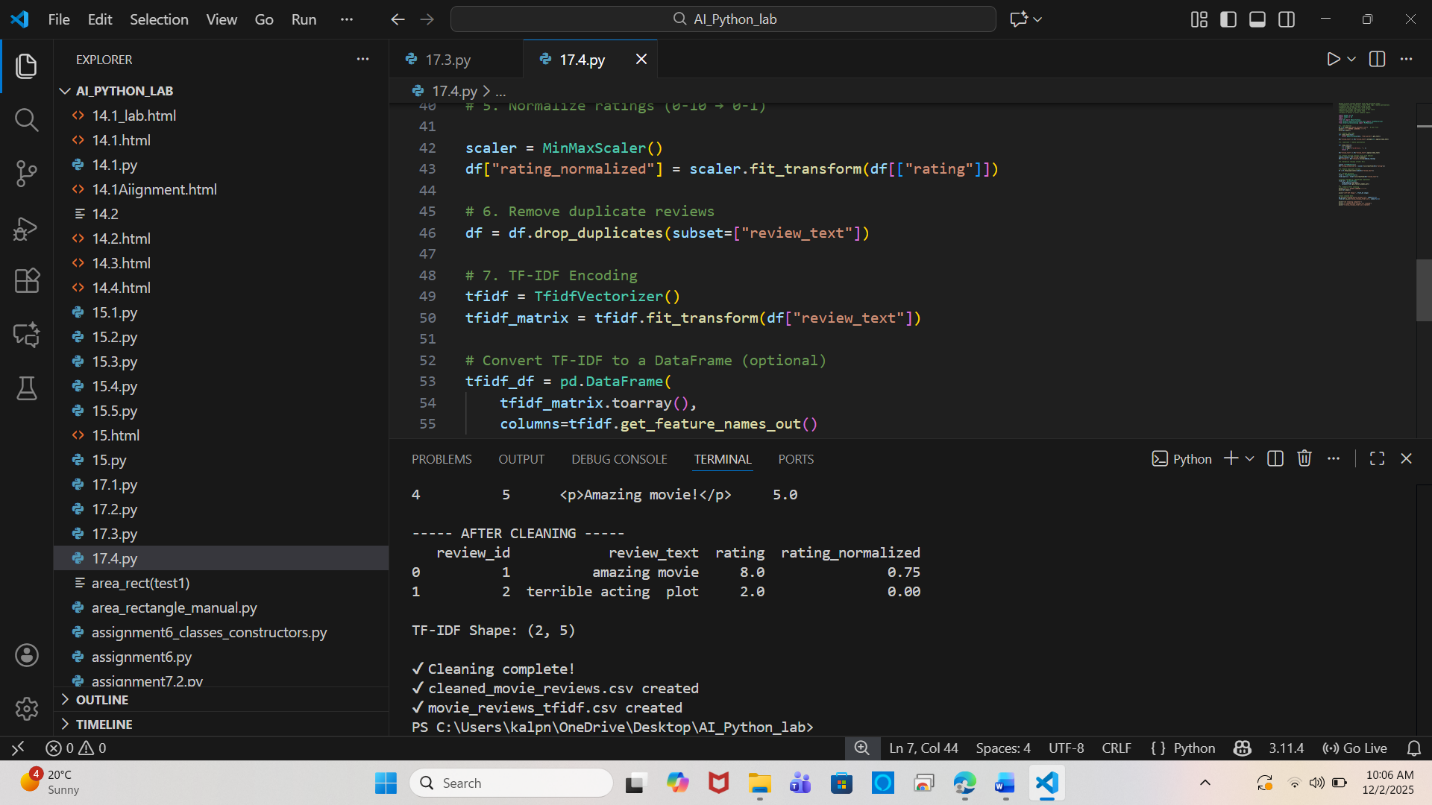
**- Generate a before vs after summary report.**

****

****

****

**Output:**

****