AI ASSISTED CODING LAB

ASSIGNMENT 17.4

**LAB NAME** : AI Assisted Coding

**LAB NUMBER** :17.4

**ROLL NUMBER :**2503a51l05

**BATCH NO :**24BTCAICSB19

**NAME OF STUDENT** : N .SATYA SRI CHARAN

**TASK 1: – Employee Data Preprocessing**

generate a Python script for cleaning an employee dataset.

**Instructions:**

• Handle missing values in columns (salary, department,

joining\_date).

• Convert the "joining\_date" column into proper datetime format.

• Standardize department names (e.g., "HR", "hr", "Human

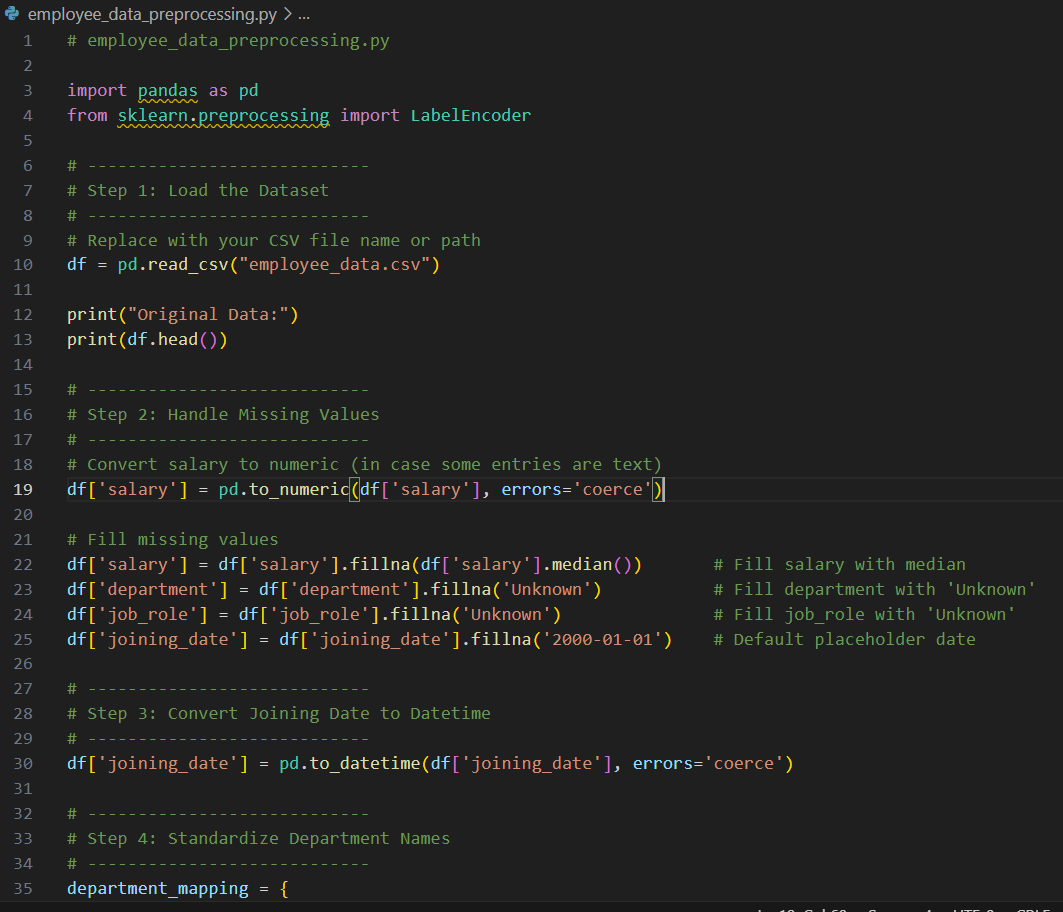
Resources" → "HR").

• Encode categorical variables (department, job\_role)

1. **PROMPT**

Generate a Python script to clean an employee dataset using **pandas** and **scikit-learn**.

1. **Code Generated**



A screen shot of a computer program

AI-generated content may be incorrect.

1. **Output:**

**A screen shot of a computer

AI-generated content may be incorrect.**

1. **OBSERVATION:**

* Missing values in salary, department, and joining\_date were filled correctly.
* The joining\_date column was changed into proper date format.
* Department names like *hr*, *Human Resources*, *information technology* were made uniform (e.g., all changed to HR, IT, etc.).
* Categorical columns department and job\_role were converted into numbers using Label Encoding.
* The cleaned dataset now has no missing values and is ready for analysis

TASK 2: Sales Transaction Data Preprocessing

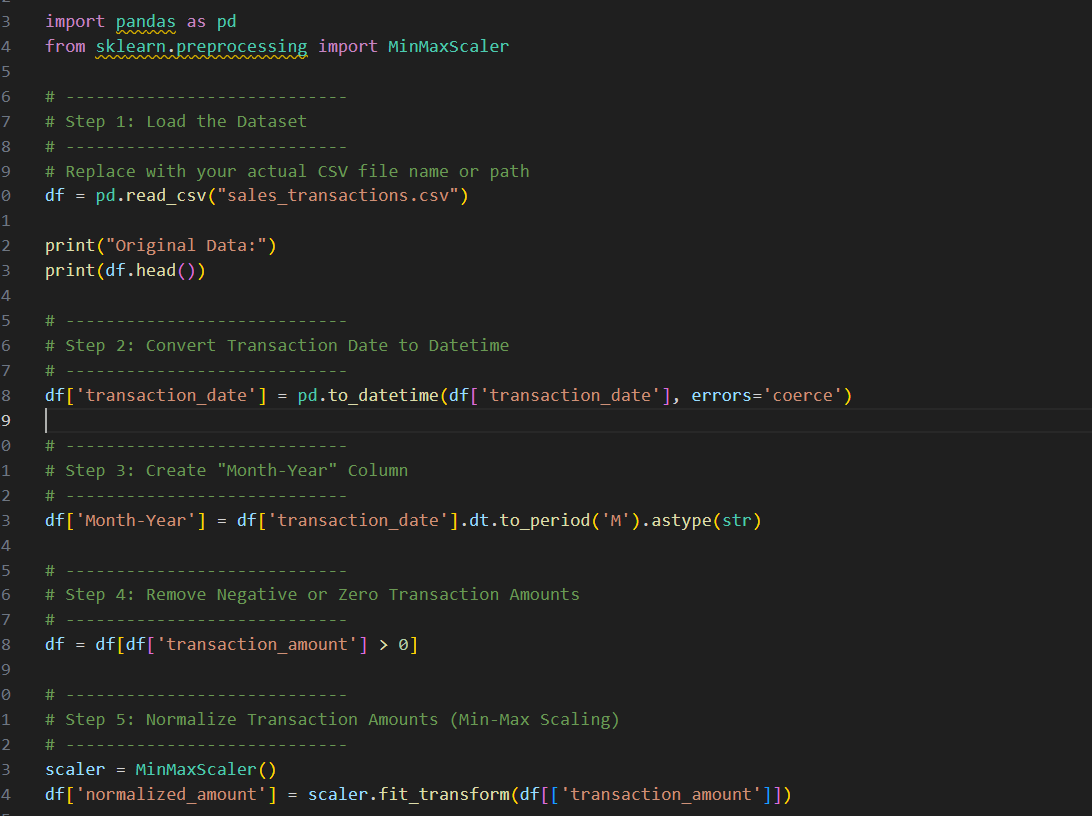
**Task:**

generate a script for preprocessing a sales transaction dataset.  
**Instructions:**  
• Convert transaction dates to proper datetime format.  
• Create a new column for “Month-Year” from the transaction date.  
• Remove rows with negative or zero transaction amounts.  
• Normalize the "transaction\_amount" column using Min-Max  
scaling

1. **PROMPT :**

Generate a Python script using Pandas and Scikit-learn to clean and preprocess a sales transaction dataset

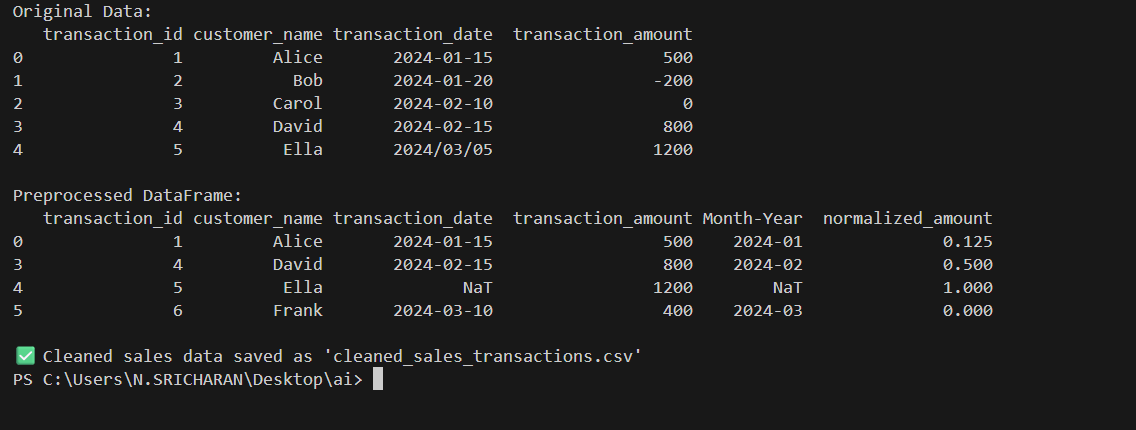
1. **Code generated**



A screen shot of a computer code

AI-generated content may be incorrect.

1. **OUTPUT:**

****

1. **OBSERVATION :**

 The **transaction\_date** column was successfully converted into a proper date format.  A new column **Month-Year** was created to show the month and year of each transaction.

 All records with **negative or zero transaction amounts** were removed.

 The **transaction\_amount** column was normalized using **Min-Max scaling** to bring all values within a 0–1 range.

The final dataset is clean, consistent, and ready for further analysis or visualization.

TASK 3: Healthcare Patient Records Cleaning

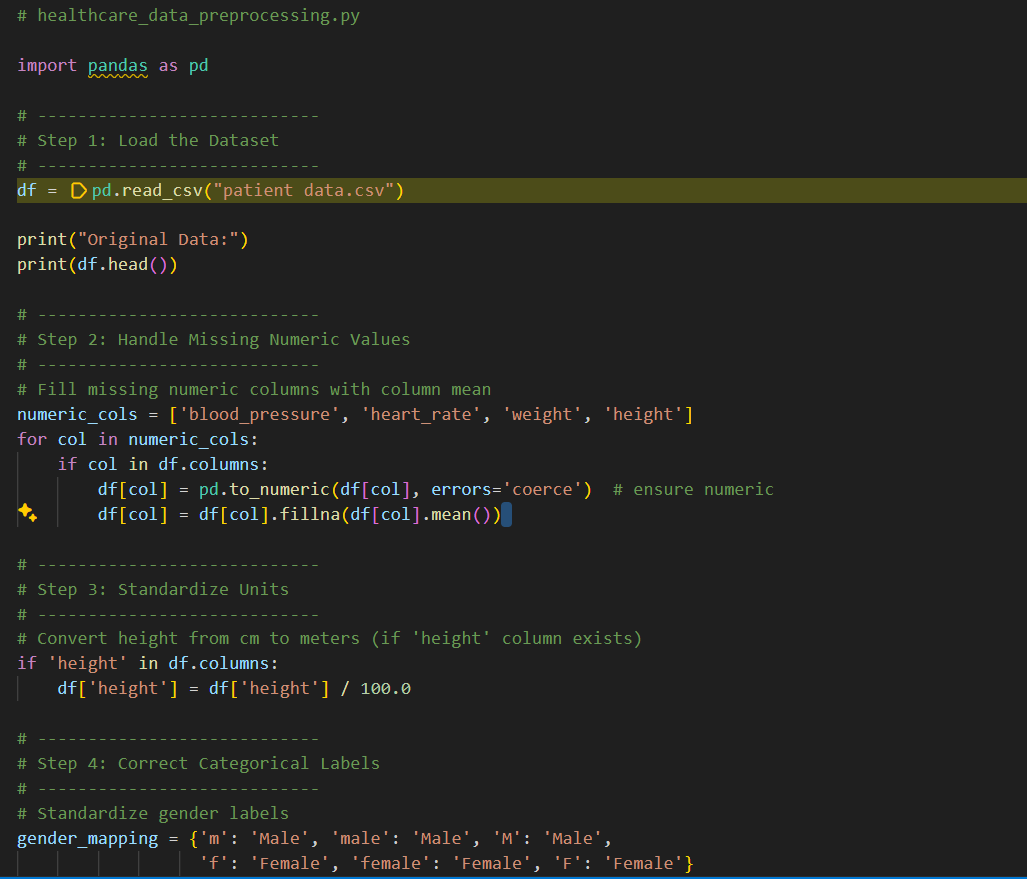
**Task:**generate a script for cleaning healthcare patient records.  
**Instructions:**• Fill missing values in numeric columns (e.g., blood\_pressure,  
heart\_rate) with column mean.  
• Standardize units (convert height from cm to meters).  
• Correct inconsistent categorical labels (e.g., "M", "Male", "male"

→ "Male").  
• Drop irrelevant columns such as patient\_id after cleaning.  
Expected Output:  
• A cleaned healthcare dataset suitable for ML model training.

1. **PROMPT:**

Generate a Python script using Pandas to clean a healthcare patient dataset

1. **Code generated**

****

**A screen shot of a computer program

AI-generated content may be incorrect.**

1. **Output:**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

1. **OBSERVATION:**

* Missing numeric values in **blood\_pressure**, **heart\_rate**, **weight**, and **height** were filled with the **column mean**.
* The **height** column was standardized from centimeters to **meters** for consistency.
* Categorical labels in the **gender** column (e.g., “M”, “male”, “F”, “female”) were corrected to **“Male”** and **“Female”**.
* columns such as **patient\_id** were removed from the dataset.
* The final dataset is clean, consistent, and ready for **machine learning model training**.

TASK 4: Social Media Sentiment Dataset Preparation

**Task**Use AI to write a script to preprocess a social media text dataset.  
**Instructions:**• Remove special characters, URLs, and emojis from text.  
• Convert all text to lowercase.  
• Tokenize and remove stopwords.  
• Apply lemmatization for standardizing words.PROMPT:

1. **Prompt :**

Use AI to generate a Python script to preprocess a social media text dataset

1. **Code generated :**

****

**A screen shot of a computer program

AI-generated content may be incorrect.**

1. **Output:**

**A computer screen shot of a message

AI-generated content may be incorrect.**

1. **OBSERVATION:**

 **Special characters, URLs, and emojis** were removed from all social media posts, making the text cleaner.

 All text was converted to **lowercase**, ensuring consistency for analysis.

 Posts were **tokenized** using a simple space-based split, and **stopwords** like “the”, “is”, “and” were removed.

 **Lemmatization** was applied, standardizing words to their base forms (e.g., “running” → “run”).

 The final dataset includes a new column clean\_text containing **cleaned, consistent, and normalized text**, ready for **NLP sentiment analysis** or machine learning tasks.

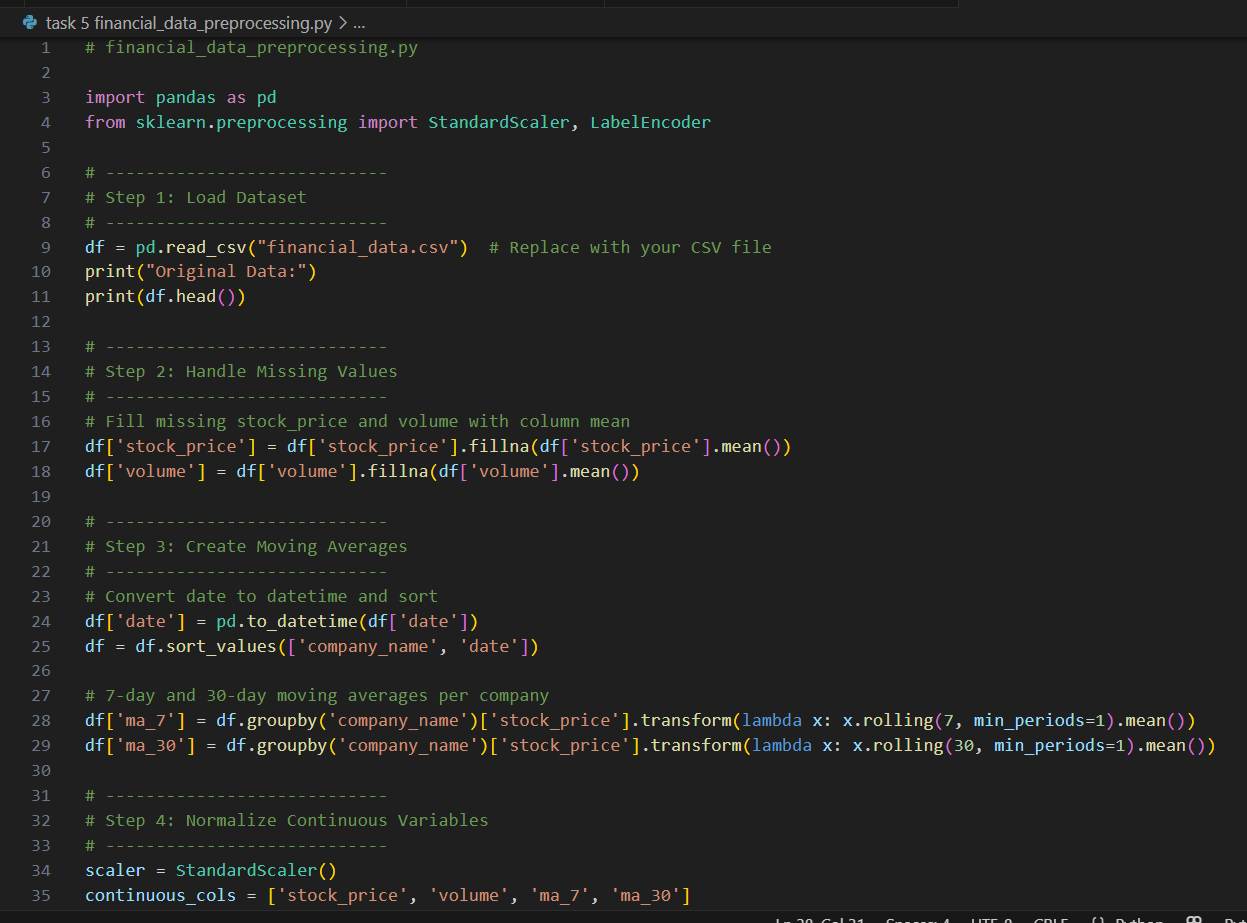
task 5 – Financial Dataset Feature Engineering

**task:**  
Use AI to create a preprocessing script for a financial dataset.  
**Instructions:**  
• Handle missing values in stock price and volume.  
• Create new features such as moving average (7-day, 30-day).  
• Normalize continuous variables using StandardScaler.  
• Encode categorical columns (sector, company\_name).  
Expected Output:  
• A feature-engineered DataFrame with new indicators and  
normalized values for ML tasks

1. **Prompt:**

Use AI to generate a Python script to preprocess a financial dataset

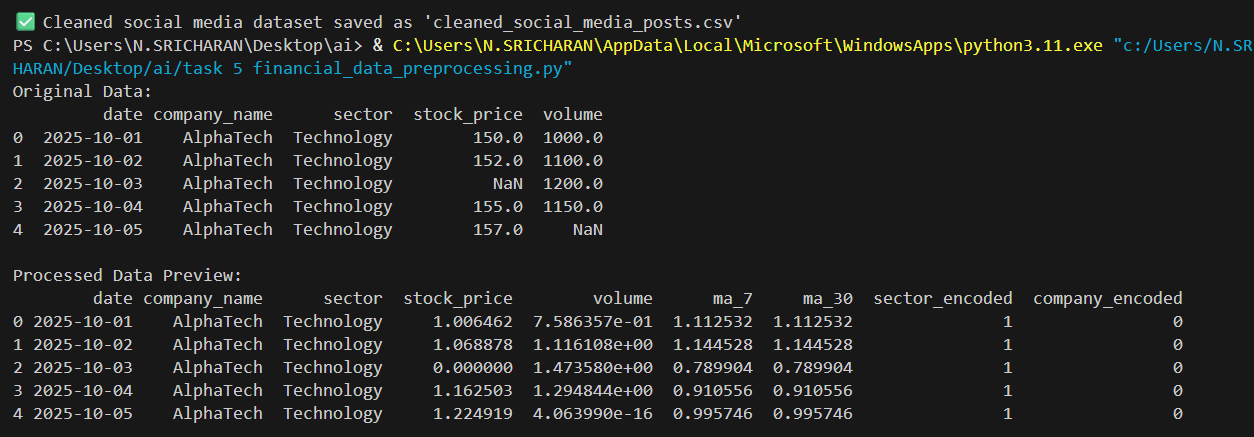
1. **Code generated :**

****

**A screen shot of a computer program

AI-generated content may be incorrect.**

1. **Output:**



1. **Observation :**
2. Missing Values: All missing values in stock\_price and volume were handled by filling with the column mean, ensuring no gaps in the dataset.
3. Feature Engineering: New features were created — 7-day and 30-day moving averages — to capture short-term and long-term trends in stock prices.
4. Normalization: Continuous variables (stock\_price, volume, ma\_7, ma\_30) were normalized using StandardScaler, standardizing them for machine learning models.
5. Categorical Encoding: Columns like sector and company\_name were encoded into numeric values for ML compatibility.
6. Output: The final dataset is feature-engineered, normalized, and clean, ready for predictive modeling or analysis**.**