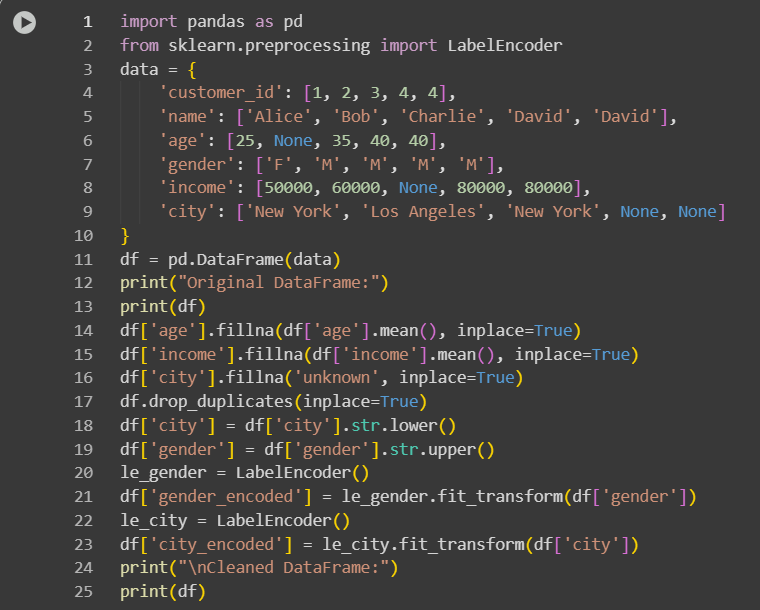
**AI LAB – 17.1**

Name: NOMULA SAI TEJA

Batch: 1

**Task 1** – Customer Data Cleaning  
Use AI to generate a Python script for cleaning a customer dataset.

Code:



Output:

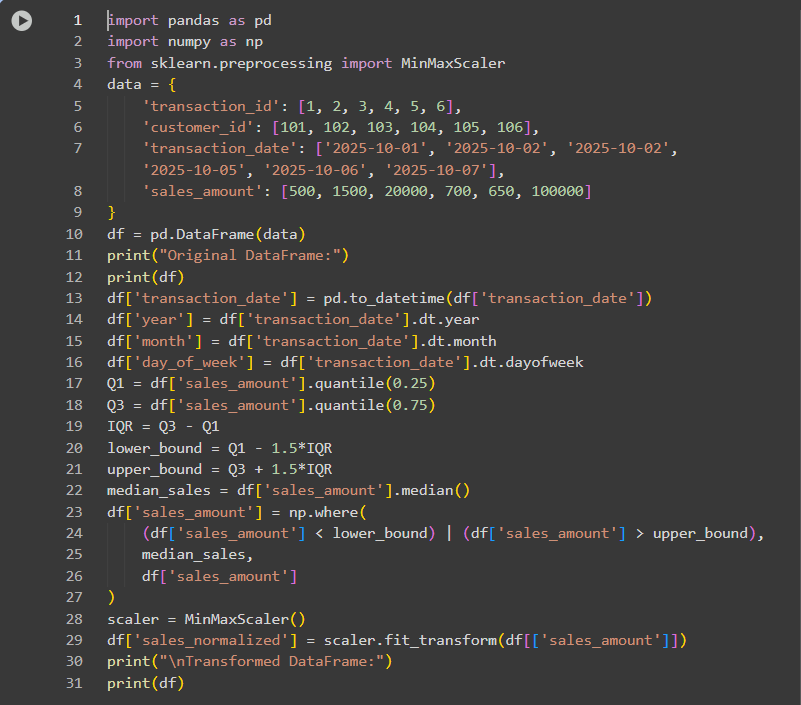


**Observation:**

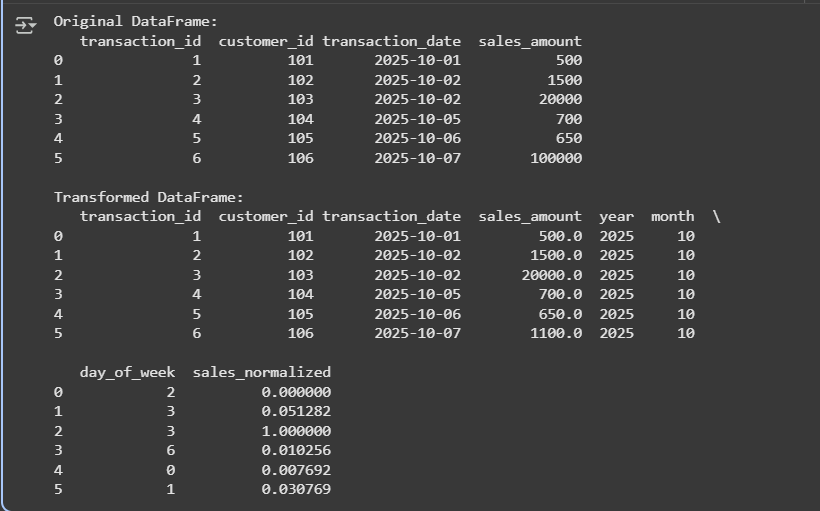
In this task, the customer dataset was cleaned to ensure it is ready for analysis. Missing values in numerical columns such as age and income were replaced with the mean values, while missing categorical values in city were filled with "unknown". Duplicate records were removed to prevent redundancy and inconsistencies. Text values, particularly city names, were standardized to lowercase, and gender values were standardized for consistency. Categorical variables like gender and city were encoded into numerical labels using LabelEncoder, enabling their use in machine learning models or statistical analysis. Overall, the cleaning process improved data quality, ensured consistency, and prepared the dataset for accurate analysis or predictive modeling.

**Task 2** – Sales Data Preprocessing  
Preprocess sales transaction data using AI assistance.

Code:



Output:



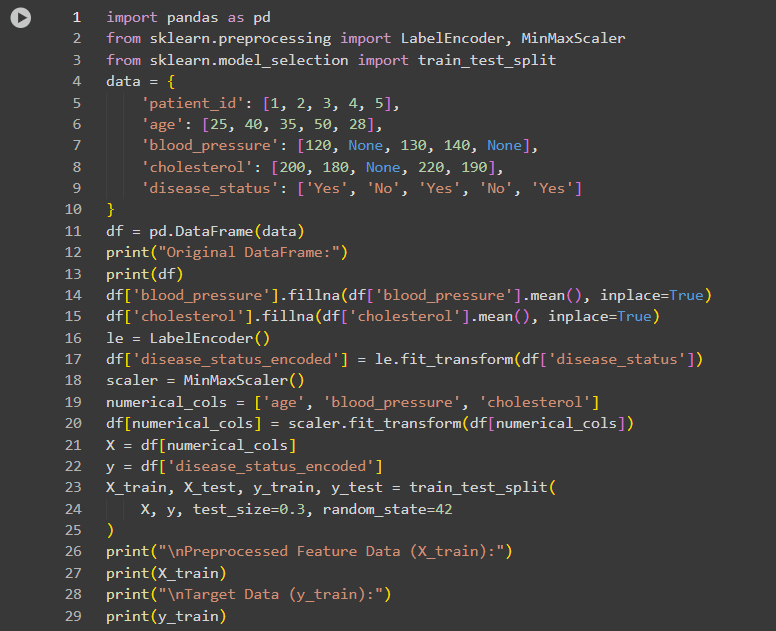
**Observation:**

In this task, the sales transaction dataset was preprocessed to prepare it for analysis and modeling. The transaction\_date column was converted into a datetime format, enabling the creation of new time-based features such as year, month, and day of the week. Outliers in the sales\_amount column were detected using the interquartile range (IQR) method and replaced with the median value to prevent skewed analysis. The sales\_amount column was also normalized using Min-Max scaling to bring all values into a 0–1 range, ensuring consistency across transactions. These preprocessing steps enhanced data quality, handled anomalies, and provided a transformed dataset with enriched features, making it suitable for accurate analysis, reporting, or predictive modeling.

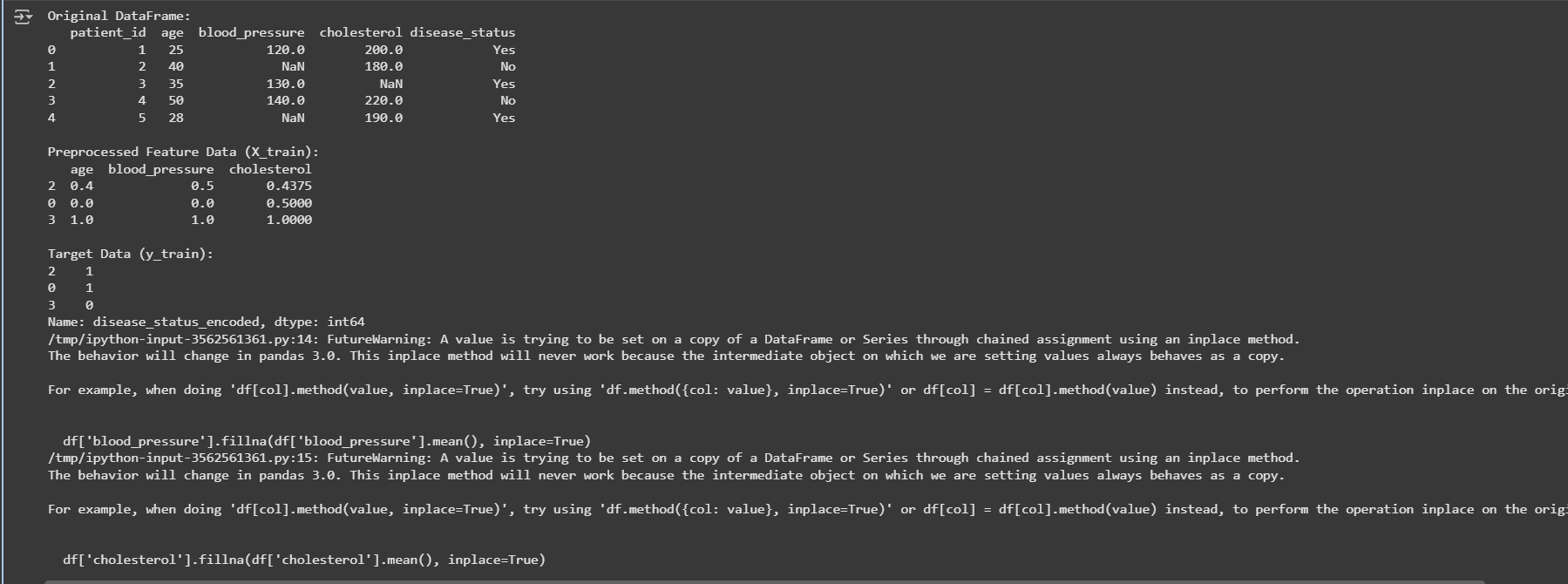
**Task 3** – Healthcare Data Preparation

Clean and preprocess a healthcare dataset using AI scripts.

Code:



Output:



**Observation:**

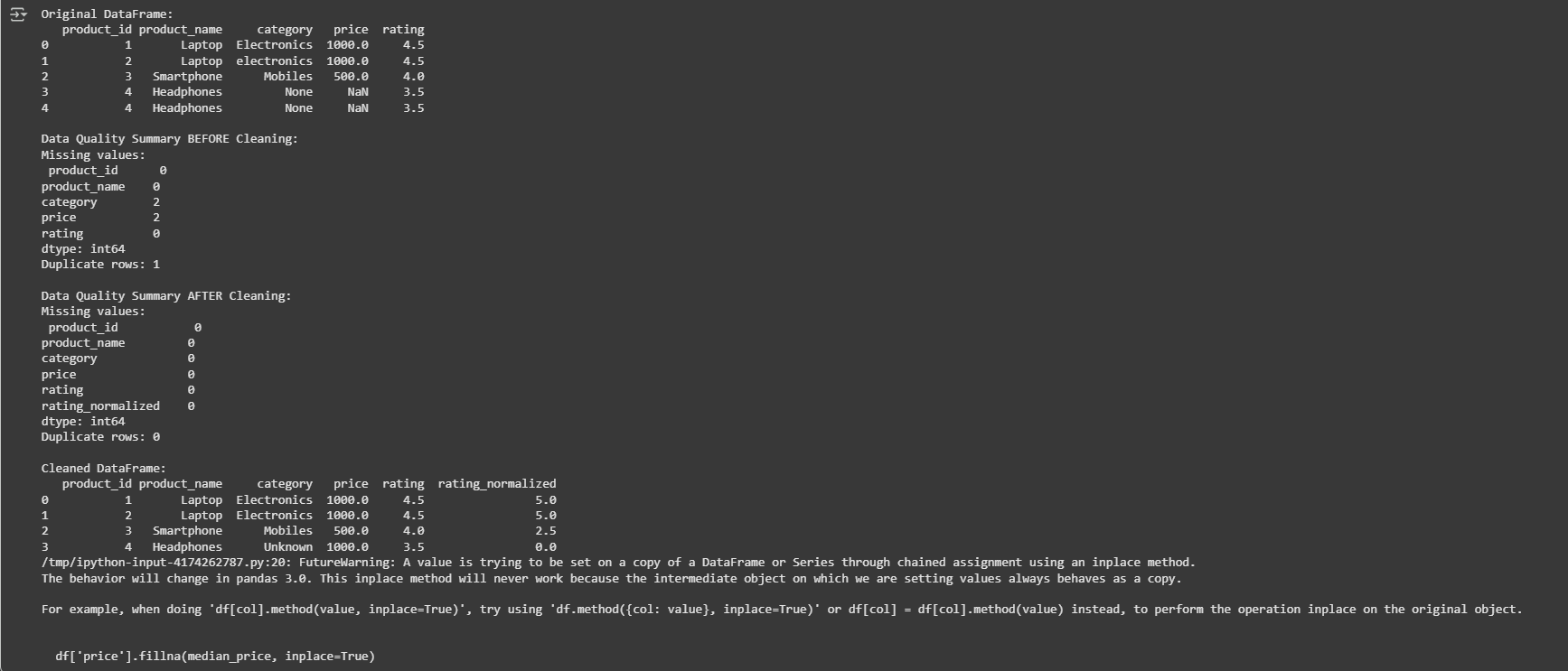
In this task, the healthcare dataset was cleaned and preprocessed to make it suitable for machine learning models. Missing values in critical numerical columns such as blood\_pressure and cholesterol were filled with the respective column means to maintain data consistency. The categorical variable disease\_status was encoded into numerical form using label encoding, converting “Yes” and “No” into 1 and 0, respectively. Numerical features including age, blood\_pressure, and cholesterol were scaled using Min-Max normalization to bring all values into a 0–1 range, ensuring uniformity for modeling. Finally, the dataset was split into training and testing sets, providing a structured framework for building and evaluating predictive models. Overall, these preprocessing steps enhanced data quality, handled missing and categorical data effectively, and prepared the dataset for accurate and reliable machine learning analysis.

**Task 4** – Real-Time Application: E-commerce Product DataCleaning  
Scenario:  
An e-commerce company has product catalog data withinconsistencies.

Code:



Output:



**Observation:**

In this task, the e-commerce product dataset was cleaned and standardized to improve data consistency and usability for analytics. Missing values in the category column were filled with "Unknown" and text fields, such as product names and categories, were standardized using consistent capitalization. Missing price values were imputed with the median price to maintain numerical consistency, and duplicate product entries were removed to prevent redundancy. Product ratings were normalized to a 0–5 scale to standardize the scoring system. A data quality summary was generated to compare missing values and duplicate rows before and after cleaning, demonstrating clear improvements. Overall, these preprocessing steps enhanced the dataset’s reliability, consistency, and readiness for further analysis or modeling.