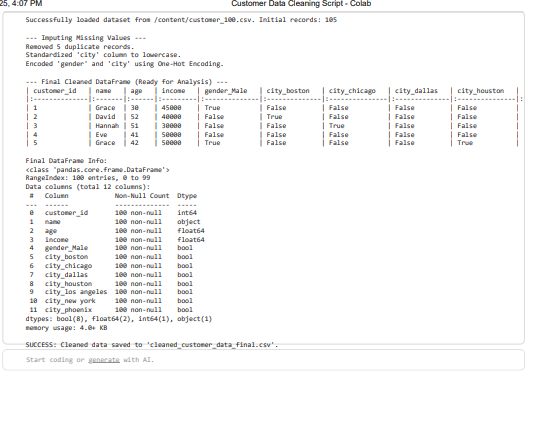
**Task 1 – Customer Data Cleaning**

Prompt: now using this file (customer\_100) generate a Python script for cleaning a customer dataset. Instructions: • Handle missing values in columns (age, income, city). • Remove duplicate records. • Standardize text values (e.g., city names in lowercase). • Encode categorical variables (gender, city). Expected Output: • A cleaned Pandas DataFrame ready for analysis

Code:



Output:

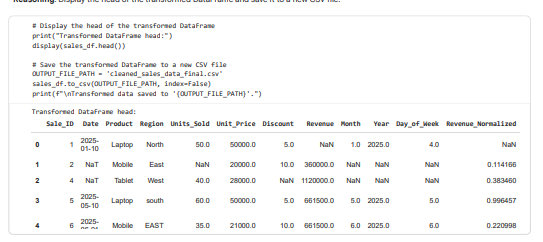
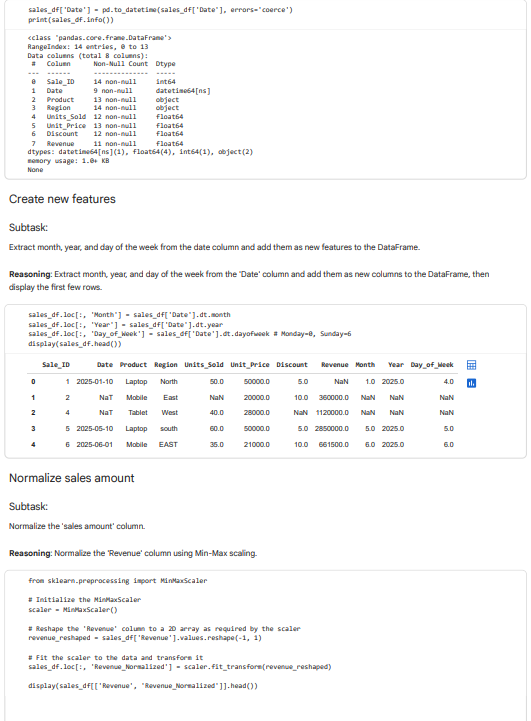
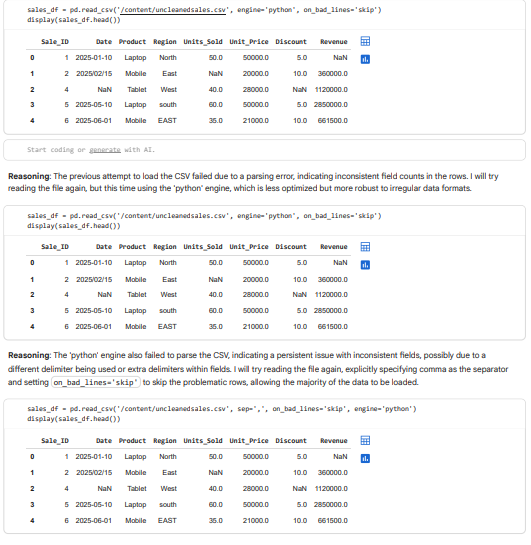


**Task 2 – Sales Data Preprocessing:**

Prompt:

Preprocess the sales transaction data from '[/content/uncleanedsales.csv](https://colab.research.google.com/drive/1Q9NlPY1v7YcW0aywj8-Bxylgf_vhpGLy)' by converting date columns to datetime format, creating new features (month, year, day of week), normalizing the sales amount column, and detecting and handling outliers in transaction amounts."

Code: and output:

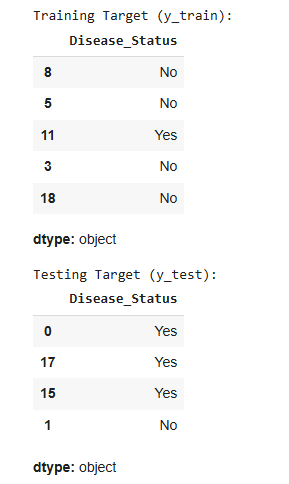
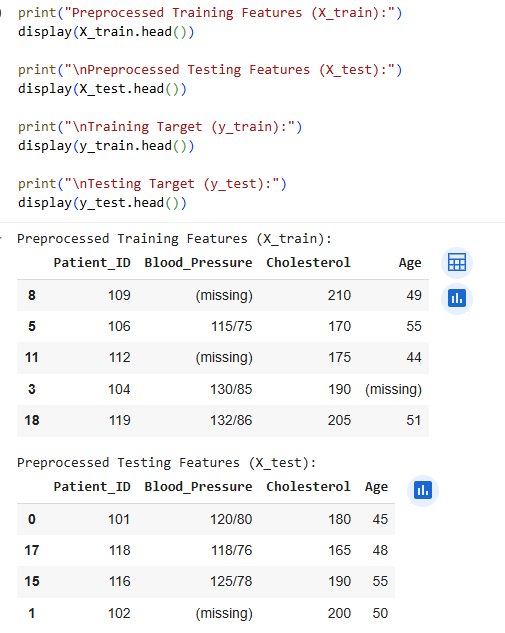
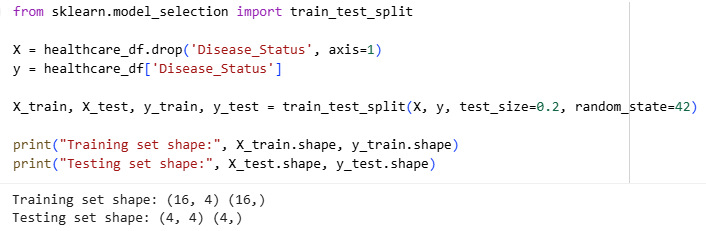
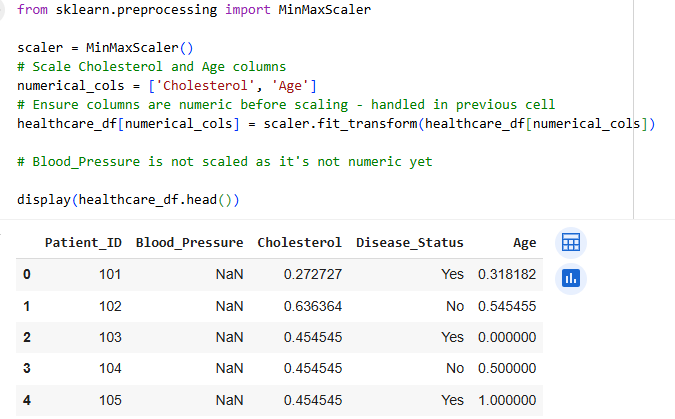
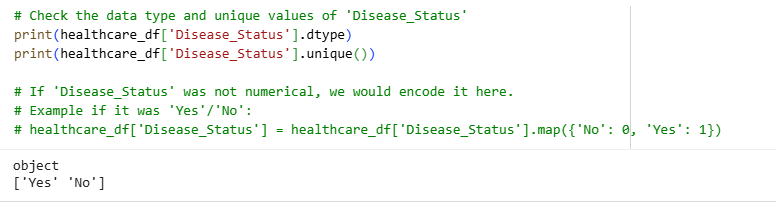
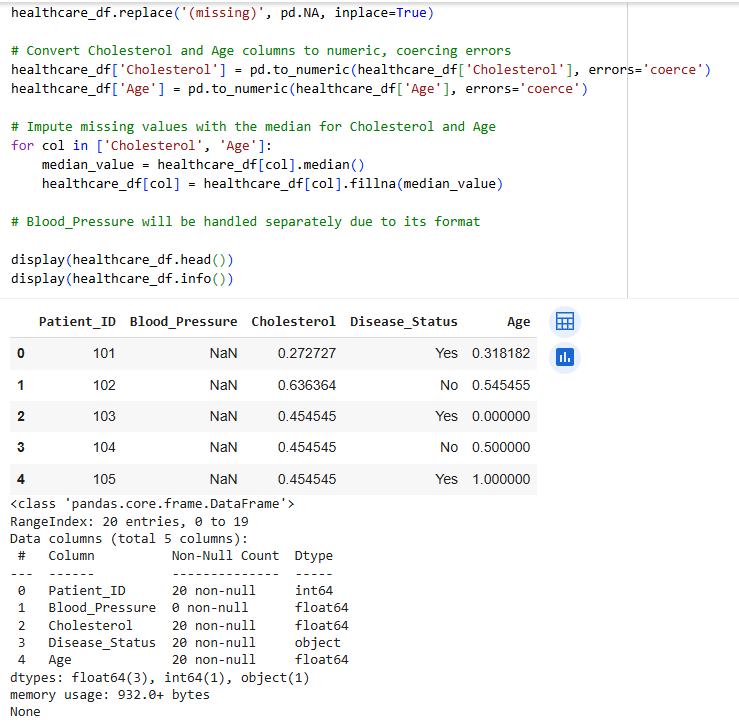
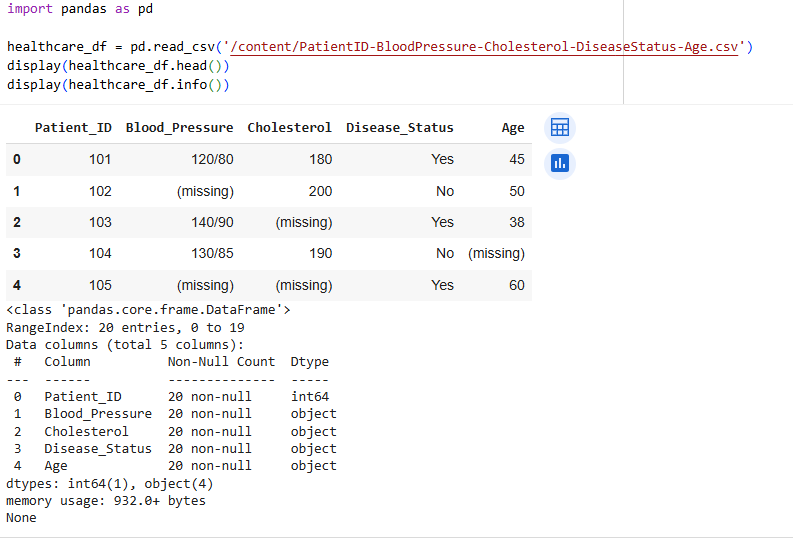


Task 3 – Healthcare Data Preparation:

Prompt:

Display the head of the training and testing datasets after preprocessing. From the given file.

Code and output :



1. Observation: **Data Loading:** We loaded the healthcare dataset from [/content/PatientID-BloodPressure-Cholesterol-DiseaseStatus-Age.csv](https://colab.research.google.com/drive/1Q9NlPY1v7YcW0aywj8-Bxylgf_vhpGLy) into a pandas DataFrame called healthcare\_df.
2. **Handling Missing Values:** We identified that the 'Blood\_Pressure', 'Cholesterol', and 'Age' columns contained '(missing)' values and were of 'object' data type. We replaced '(missing)' with pd.NA. We then converted the 'Cholesterol' and 'Age' columns to a numeric type, coercing errors for values that couldn't be converted. Finally, we imputed the missing values in these numeric 'Cholesterol' and 'Age' columns with their respective medians. We noted that 'Blood\_Pressure' requires a different approach due to its format ('systolic/diastolic').
3. **Encoding Categorical Features:** We checked the 'Disease\_Status' column and confirmed it contains 'Yes' and 'No' values of 'object' type. We noted that if necessary, this column could be encoded using techniques like Label Encoding or One-Hot Encoding, but for now, it remains as is.
4. **Scaling Numerical Features:** We applied Min-Max scaling to the 'Cholesterol' and 'Age' columns to normalize their values between 0 and 1. The 'Blood\_Pressure' column was not scaled as it's not yet in a numerical format.
5. **Splitting the Dataset:** We split the preprocessed data into features (X) and the target variable (y), which is 'Disease\_Status'. We then further split these into training (X\_train, y\_train) and testing (X\_test, y\_test) sets, with 80% of the data for training and 20% for testing.
6. **Verification:** We displayed the head of the resulting training and testing sets to confirm the structure and contents after the preprocessing steps.

In essence, we have loaded, cleaned, imputed missing values (for Cholesterol and Age), scaled numerical features (Cholesterol and Age), and split the healthcare dataset, preparing it for potential machine learning model training and evaluation.