ASSIGNMENT – 17

H.NO: 2403A52139

# Use AI to 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).

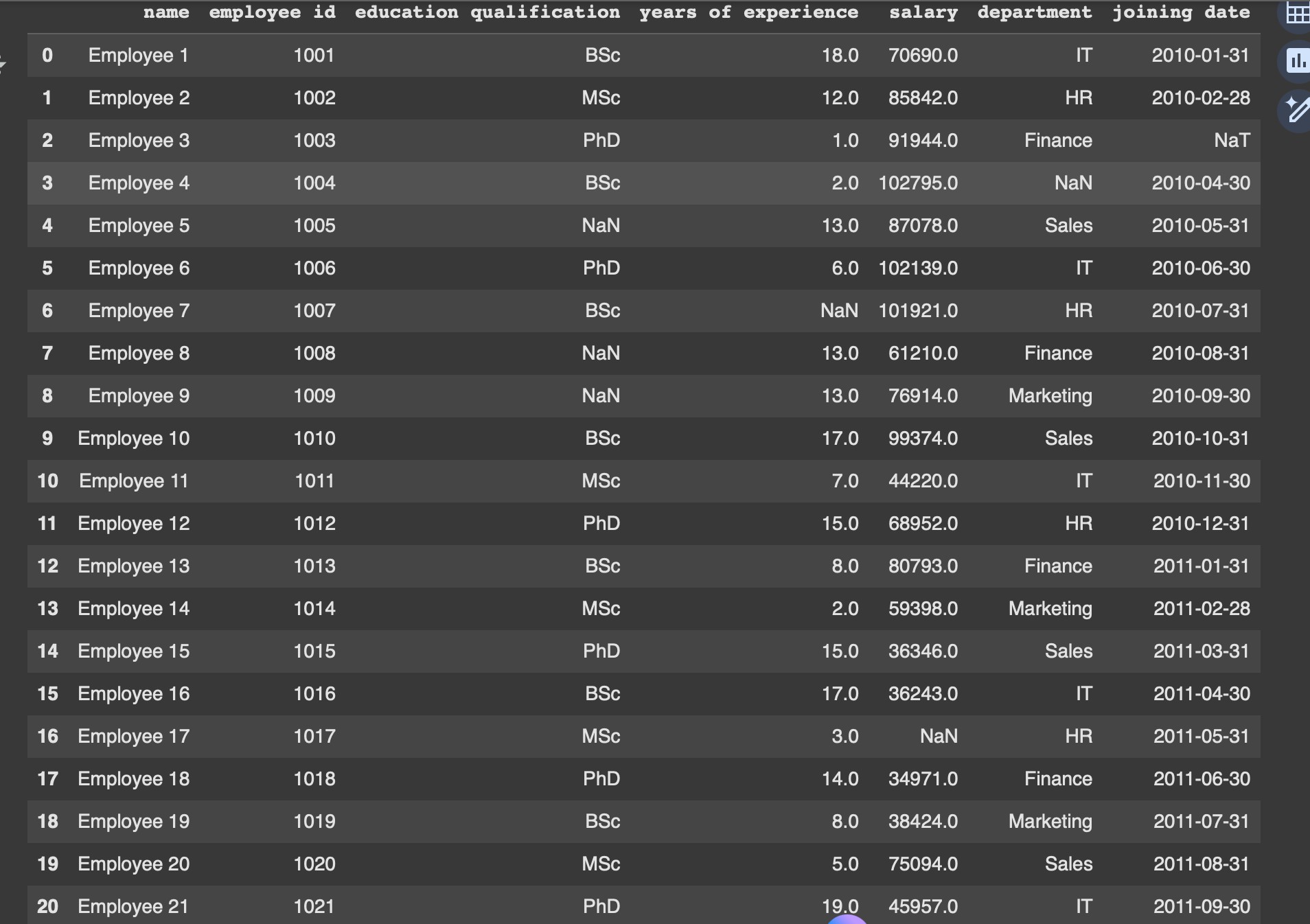
# Expected Output:

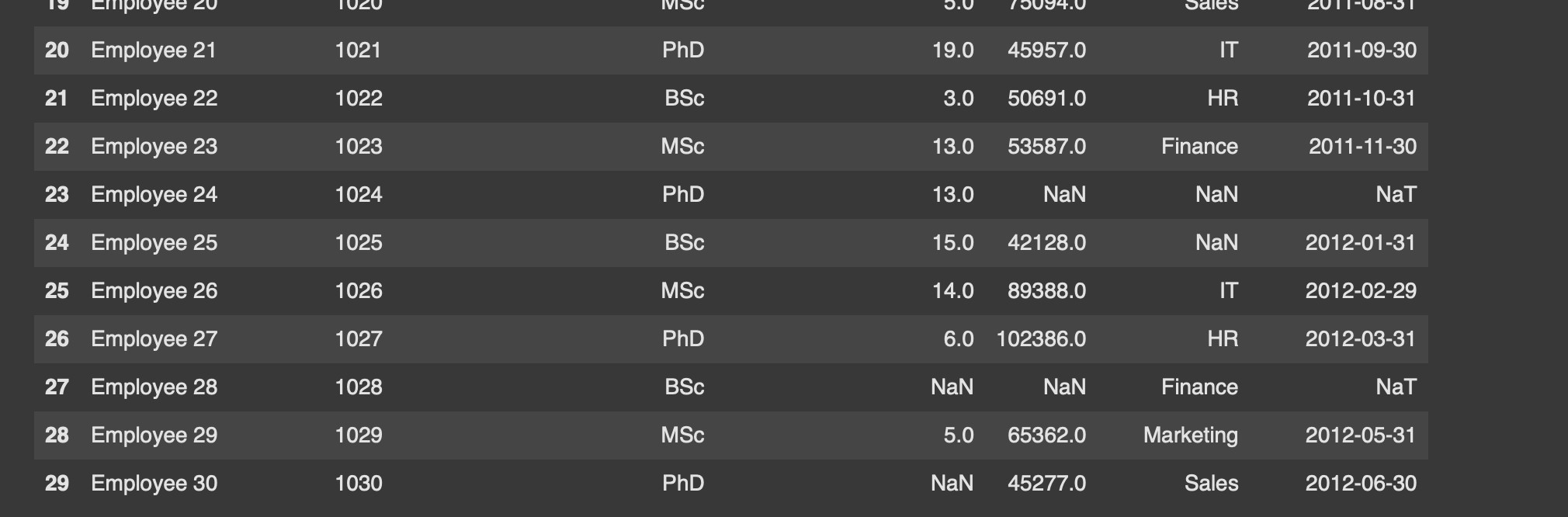
* A cleaned Pandas DataFrame with consistent departments, proper dates, and encoded features.

**Expected code :**

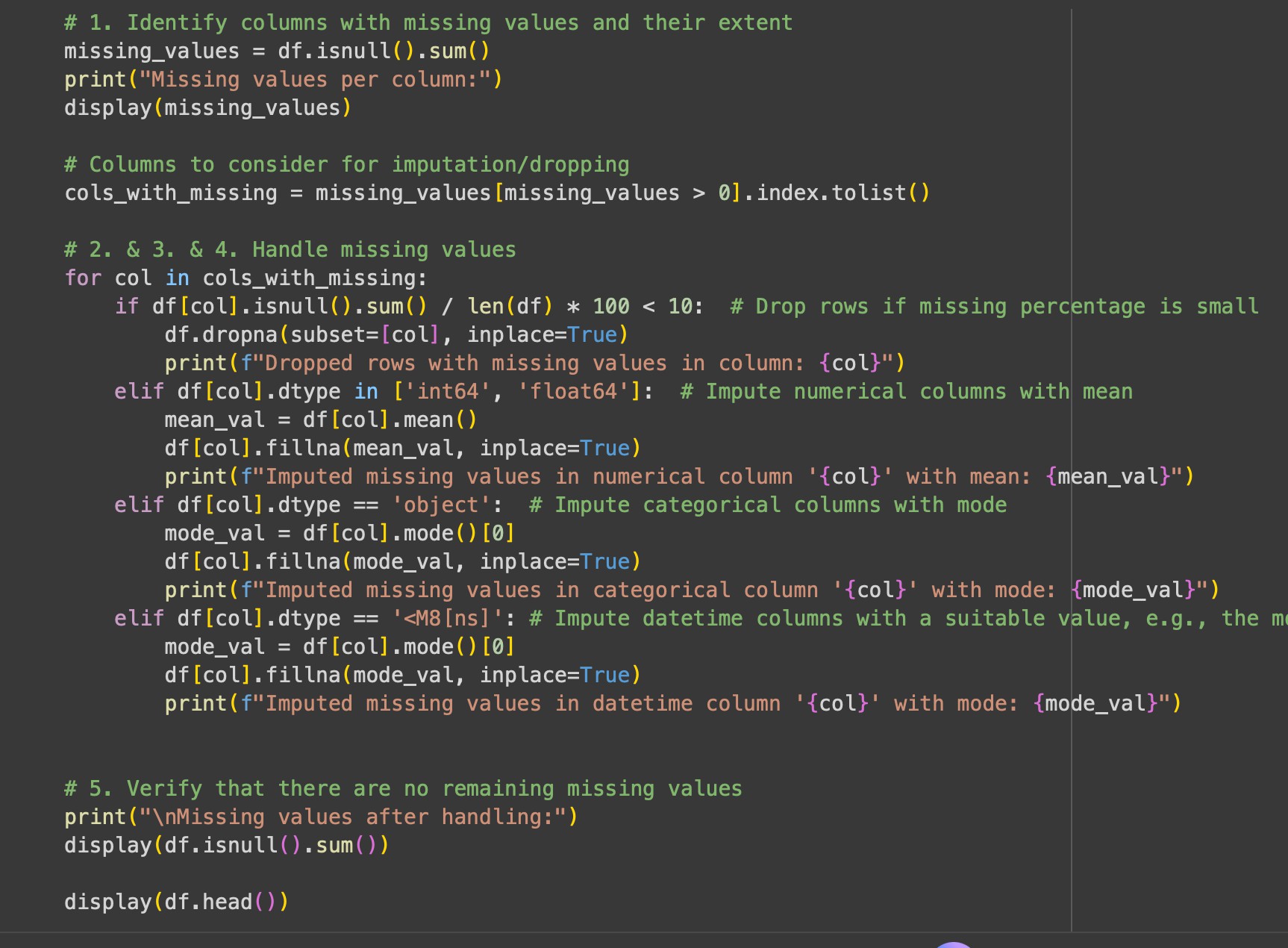
****

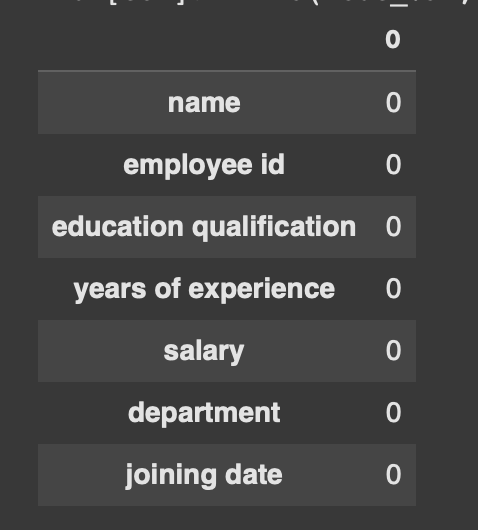
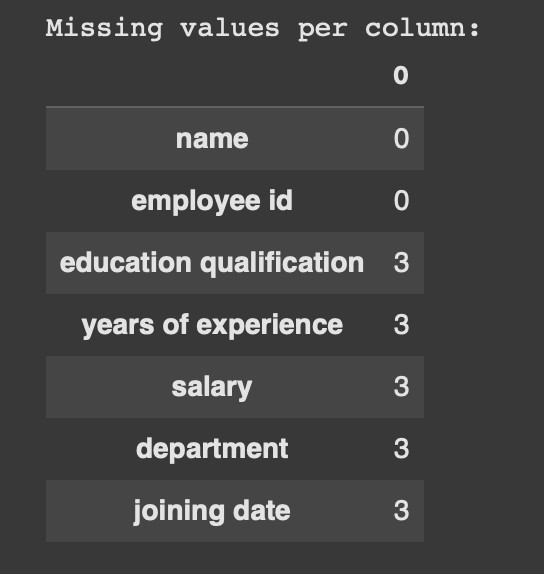
**Dataset:**

****

****

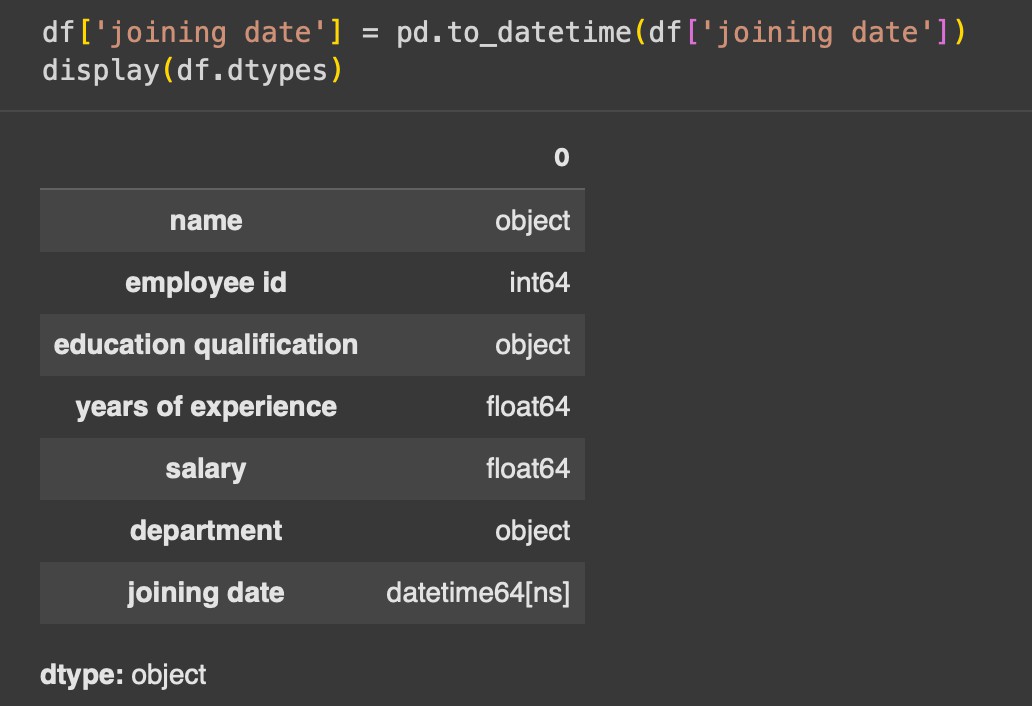
# Identifying and handling missing values:

****

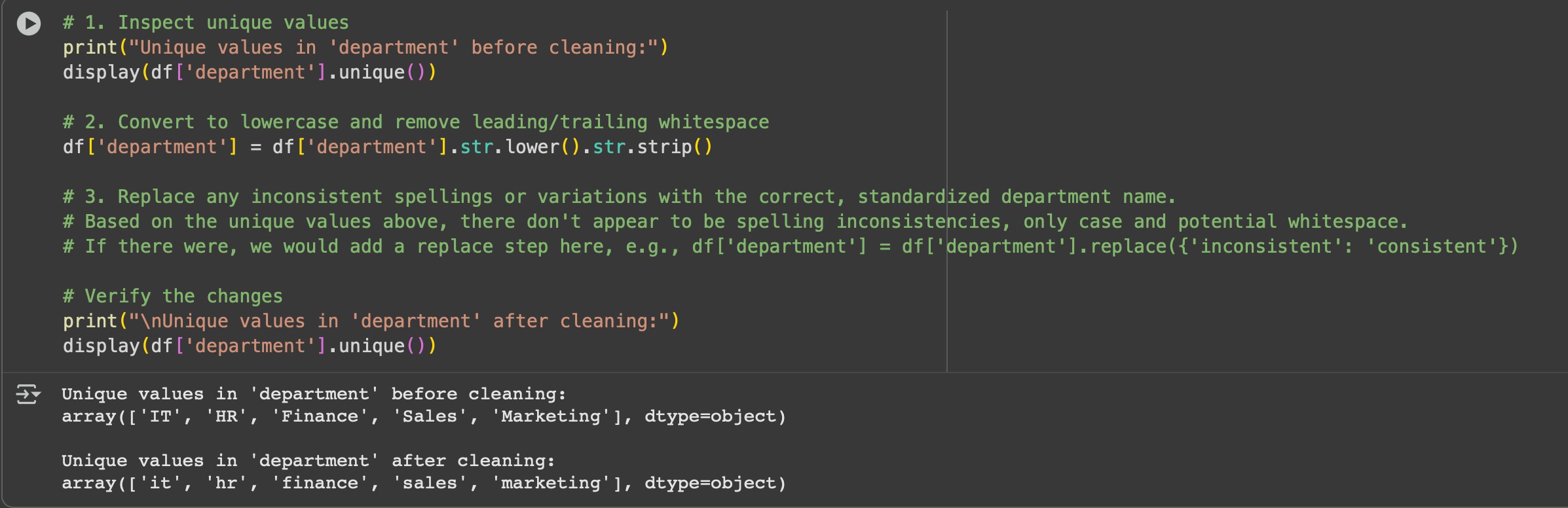
****



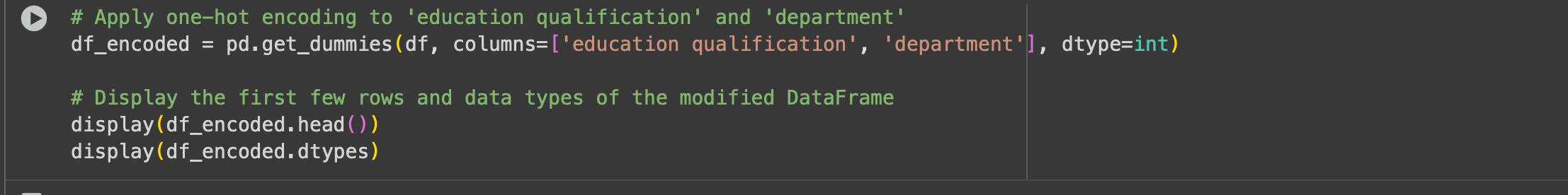
**Converting joining date to proper date format:**

****

**Standardize department names :**

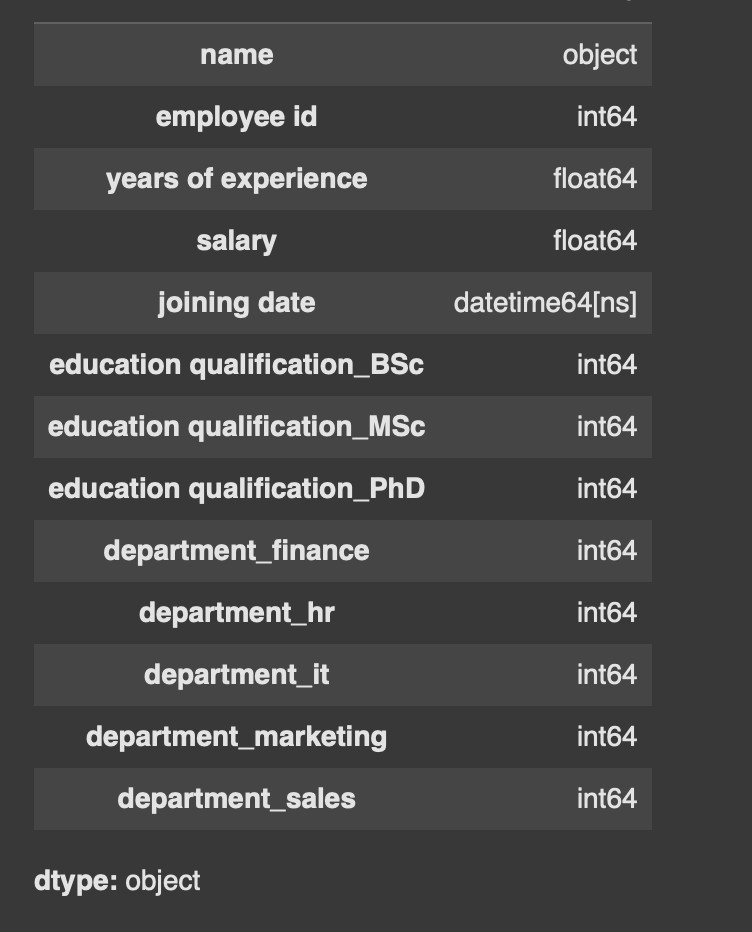
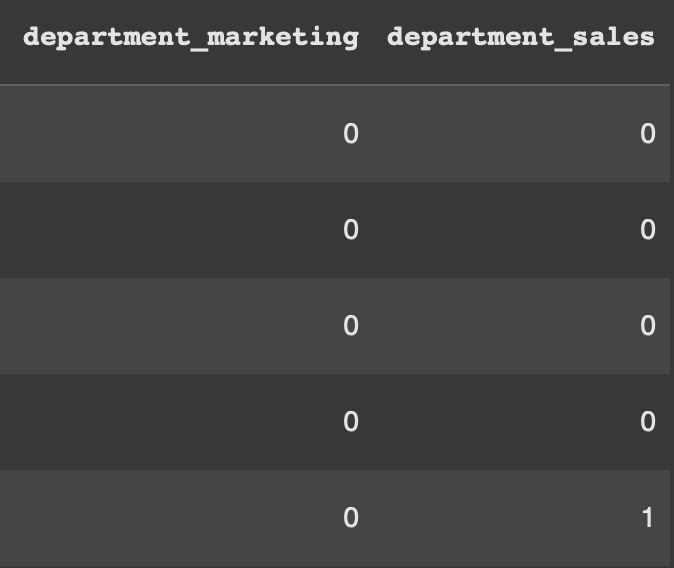
****

**Encoding data:**

****

**Expected output :**

****

****

**TASK 2:**

Use AI to generate a script for preprocessing a sales transaction dataset.

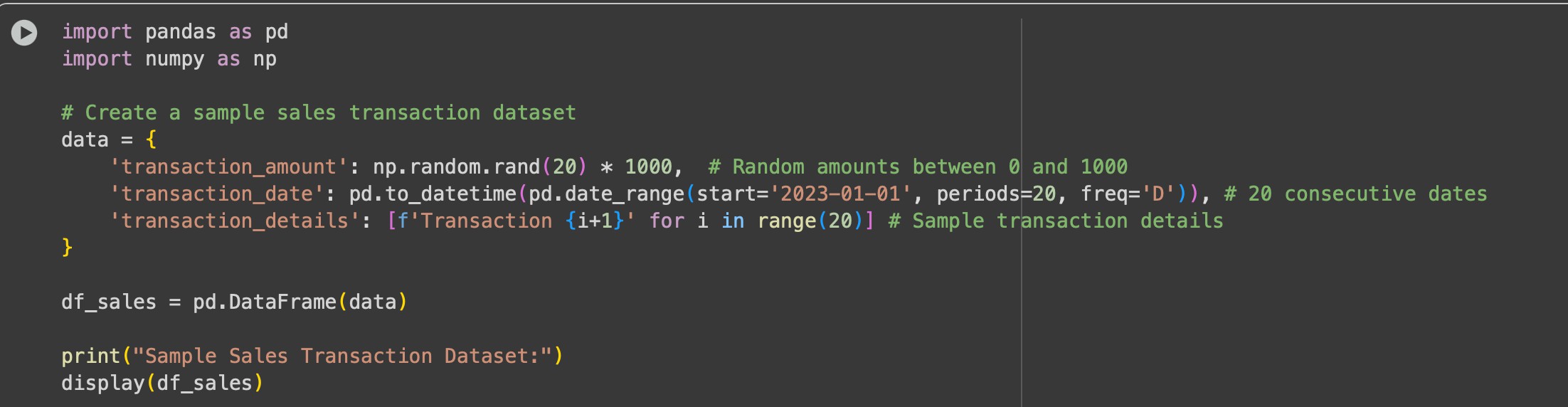
**Instructions:**

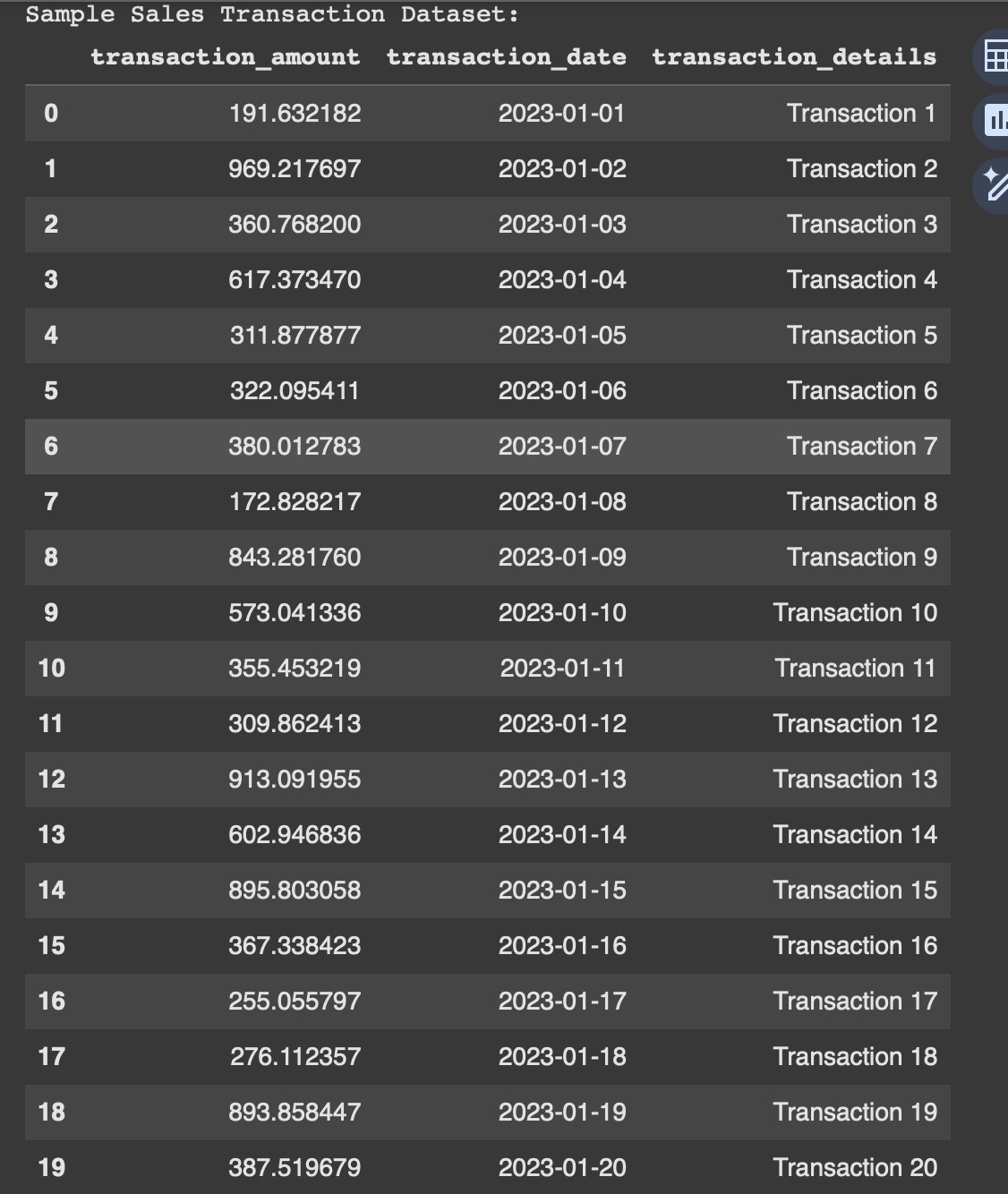
* Convert transaction dates to proper date-time format.
* Create a new column for “Month-Year” from the transaction date.
* Remove rows with negative or zero transaction amounts.
* Normalize the "transaction\_amount" column using Min-Max scaling.

**Expected Output:**

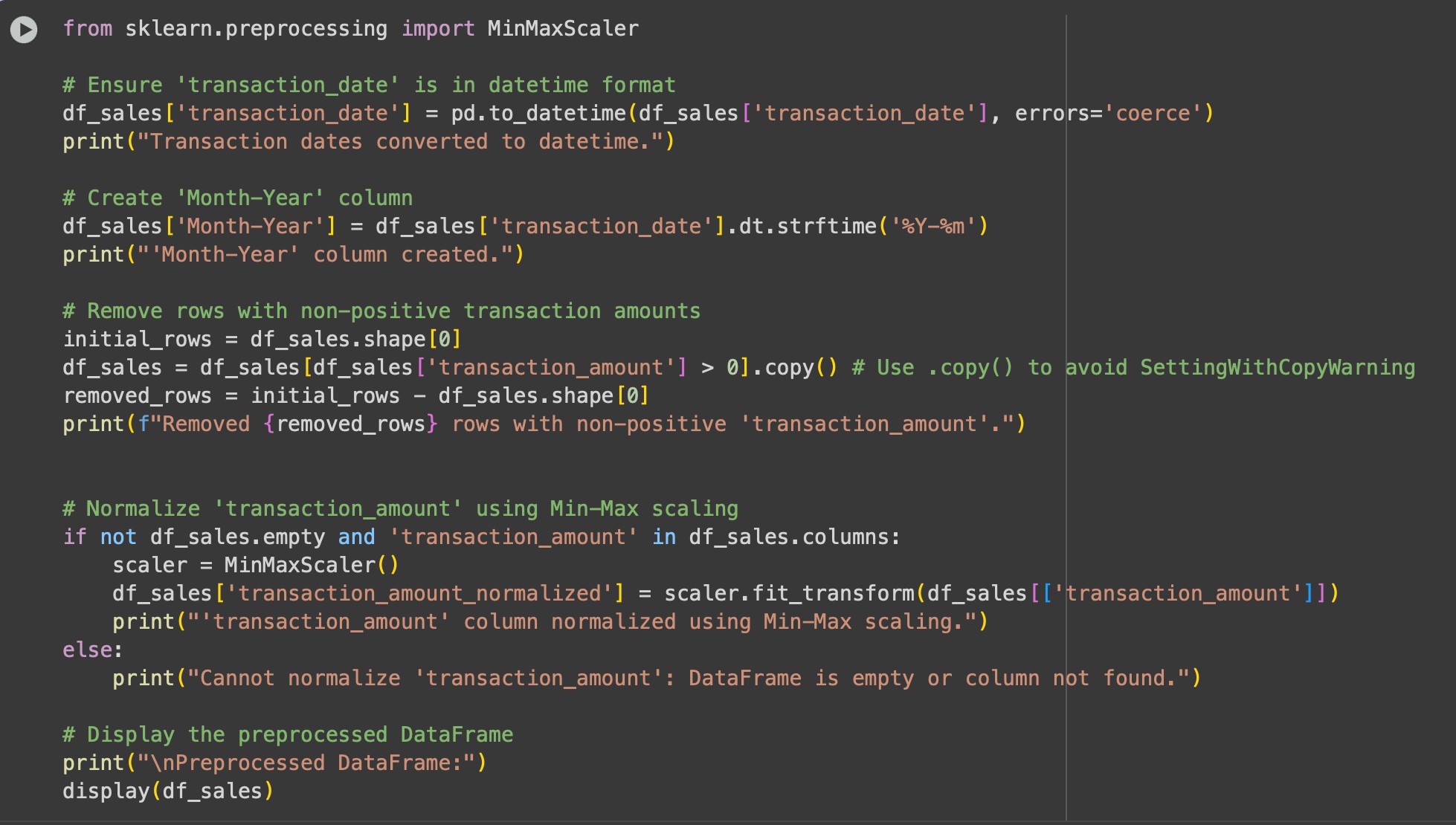
* A preprocessed DataFrame with valid dates, normalized amounts, and no invalid records.

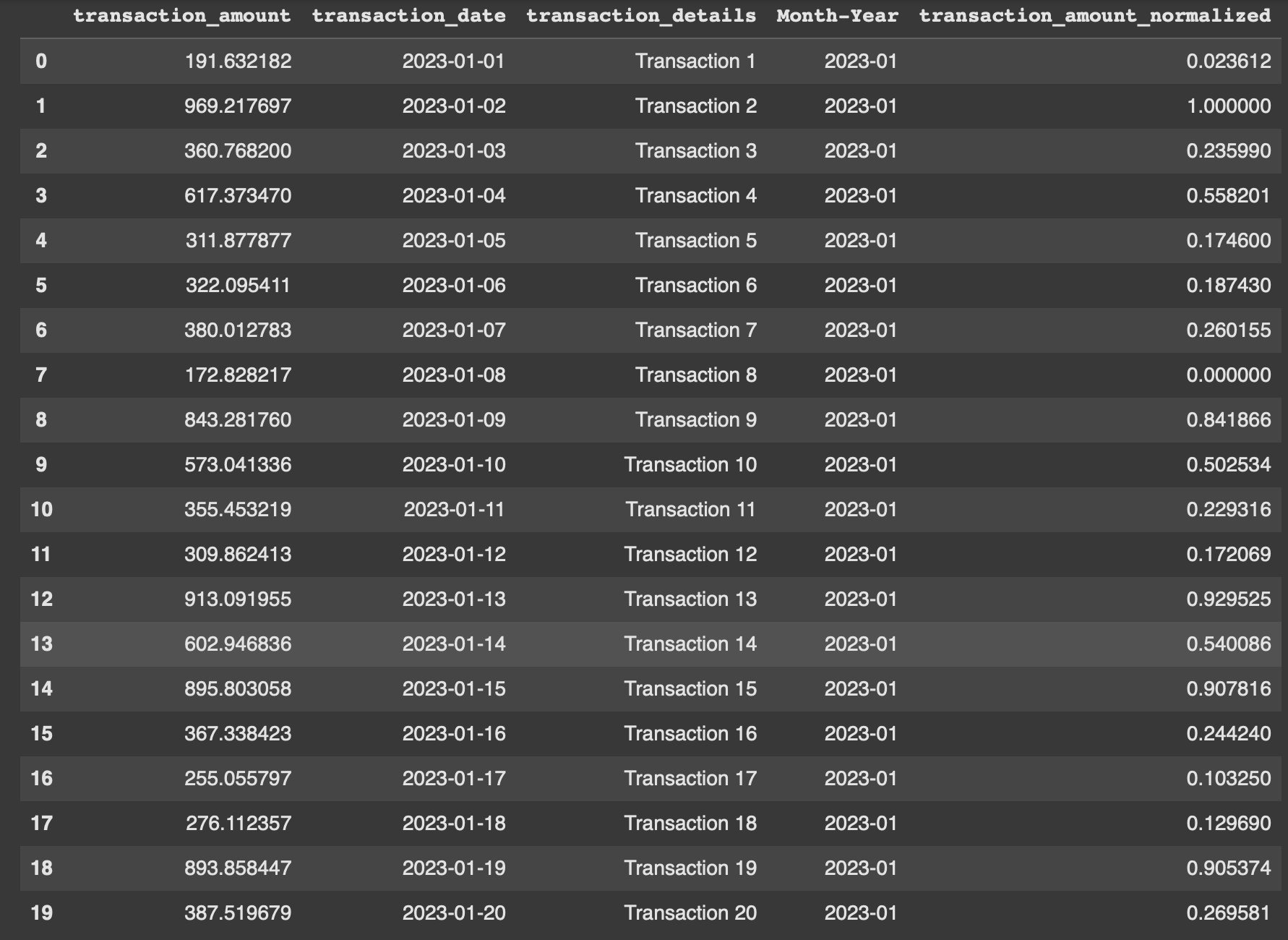
**Expected code and sample dataset :**

****

****

**Converting date format and removing irrelevant columns :**

****

****

**TASK 3:**

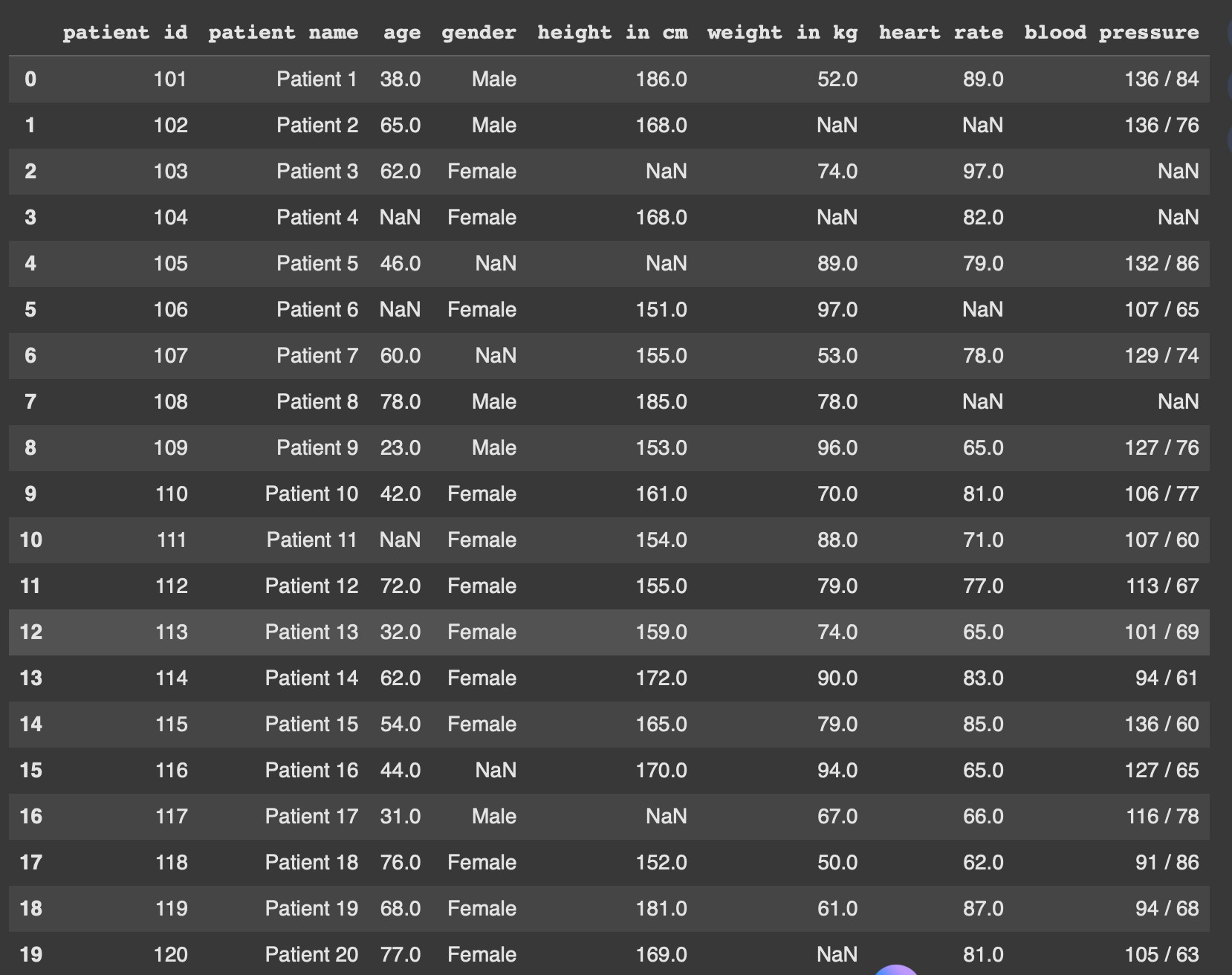
**Use AI to generate a script for cleaning healthcare patient records. Instructions:**

* 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.

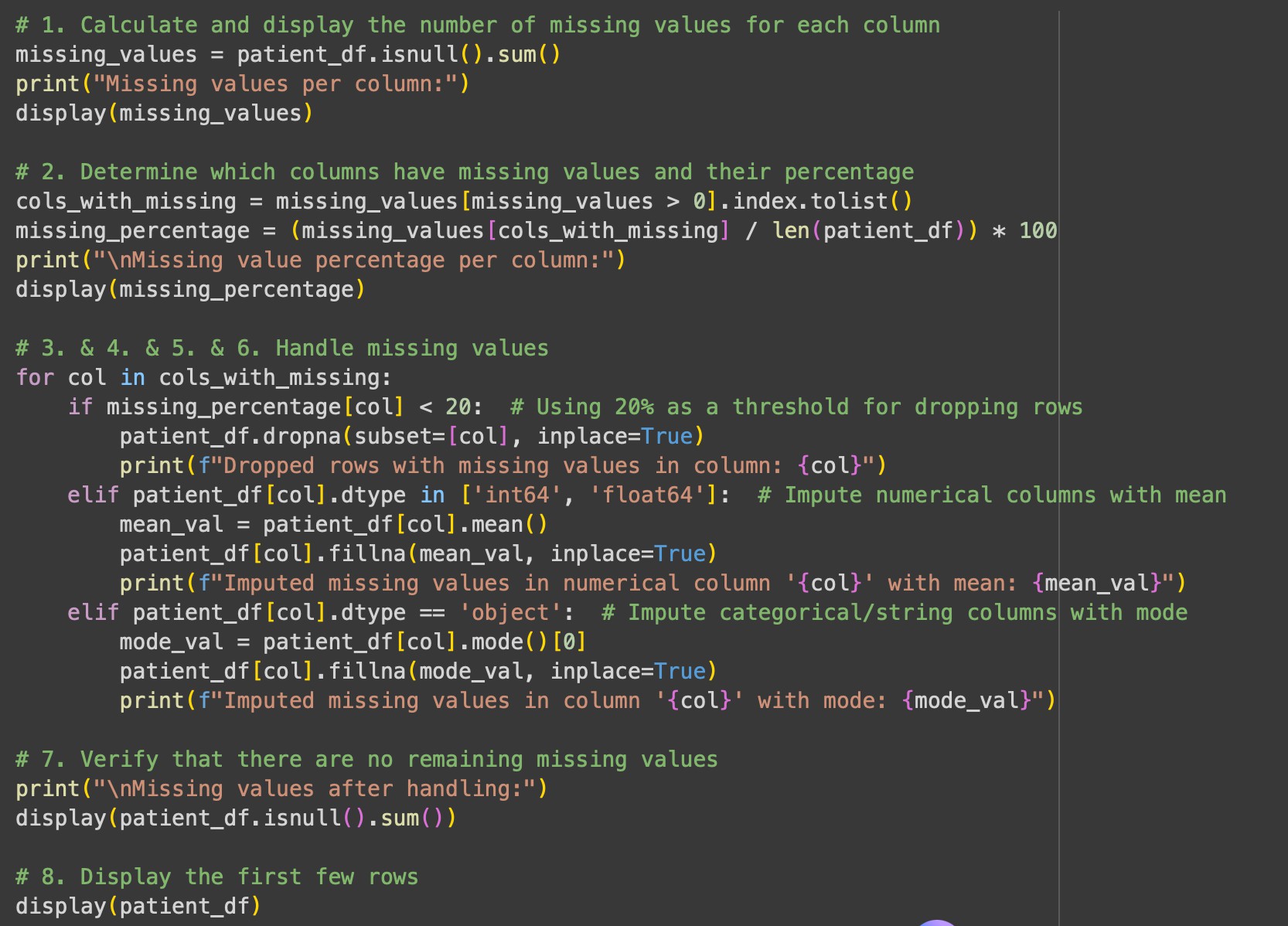
**Expected Output:**

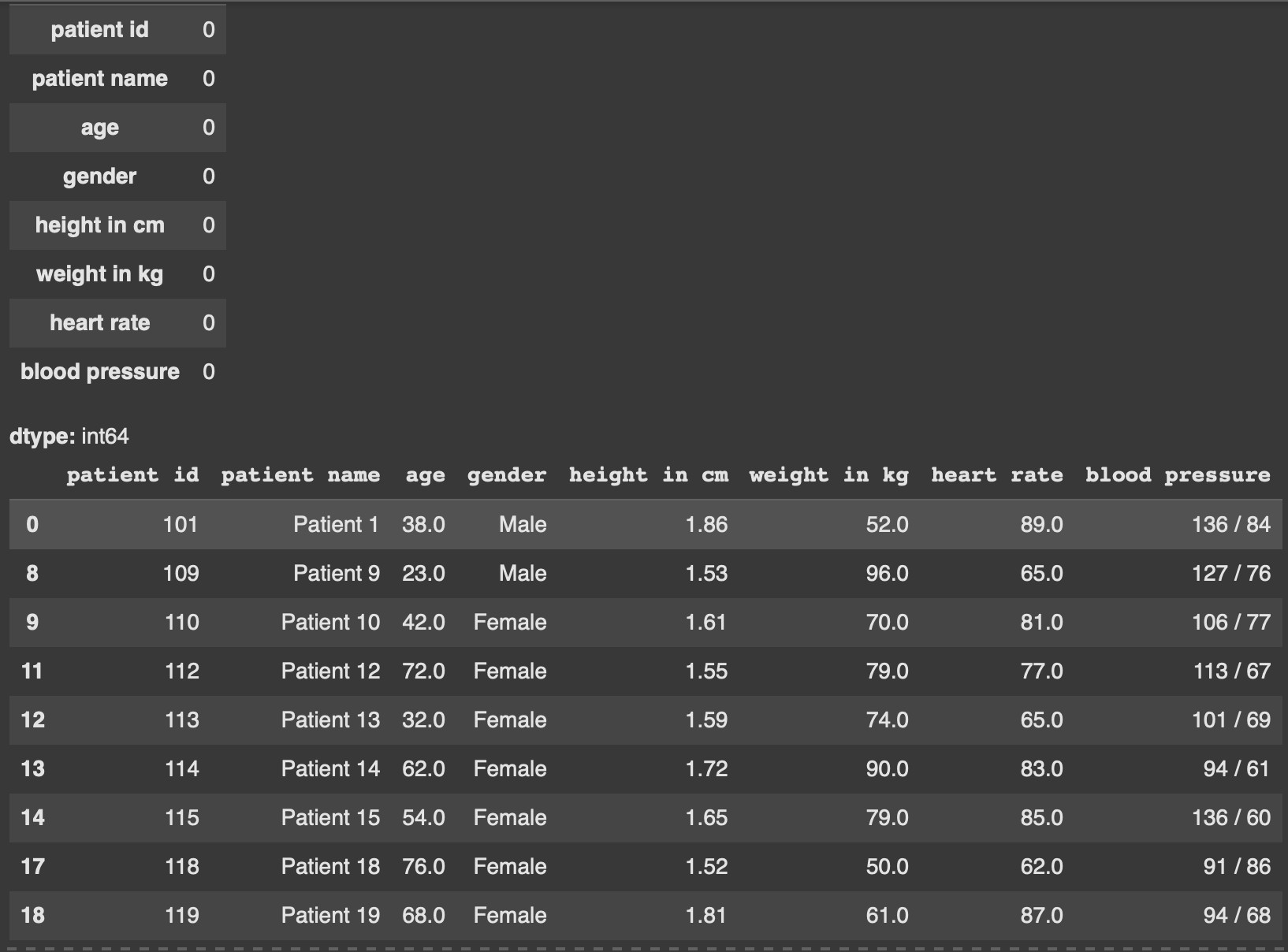
* **A cleaned healthcare dataset suitable for ML model training**

**Expected and sample dataset :**

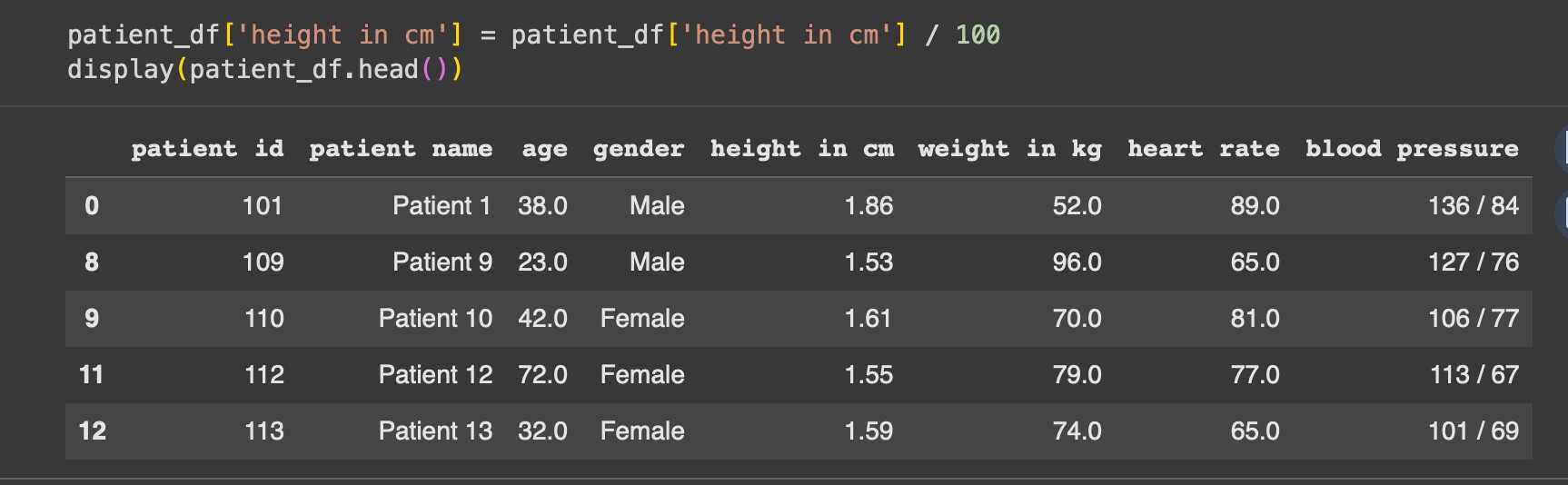
****

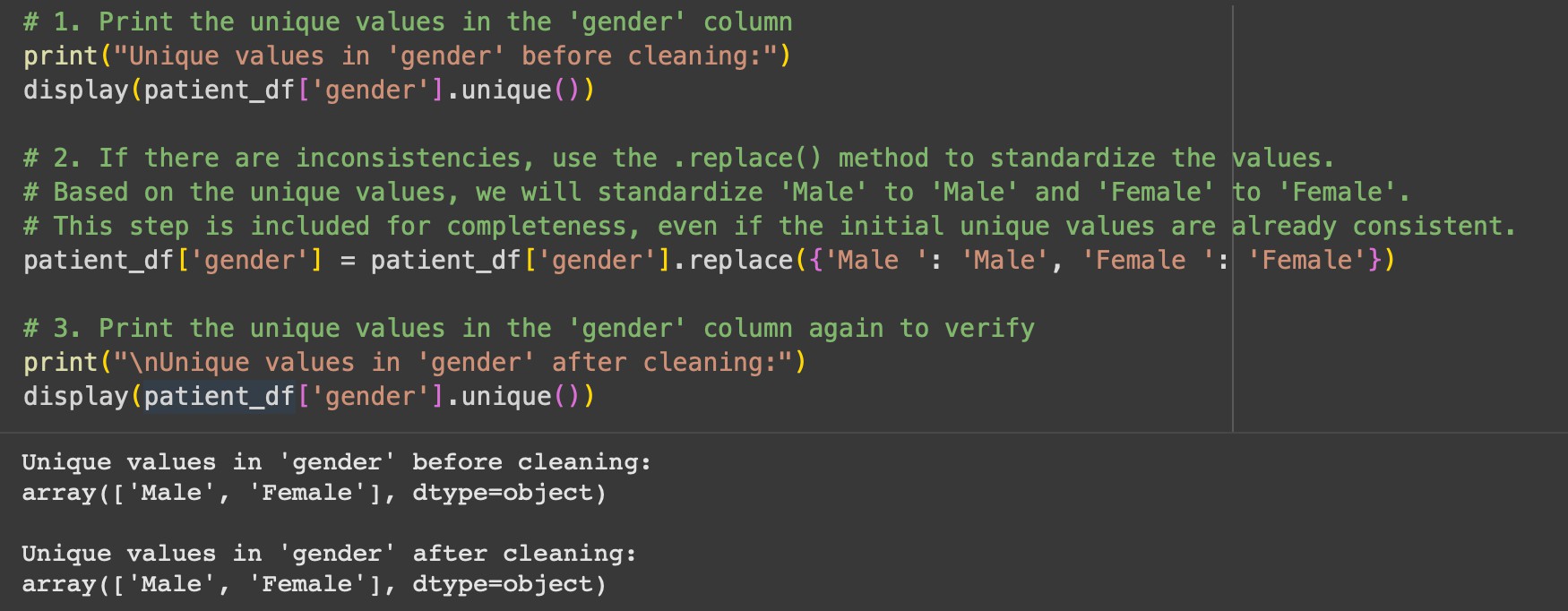
**Handling and correcting missing values:**

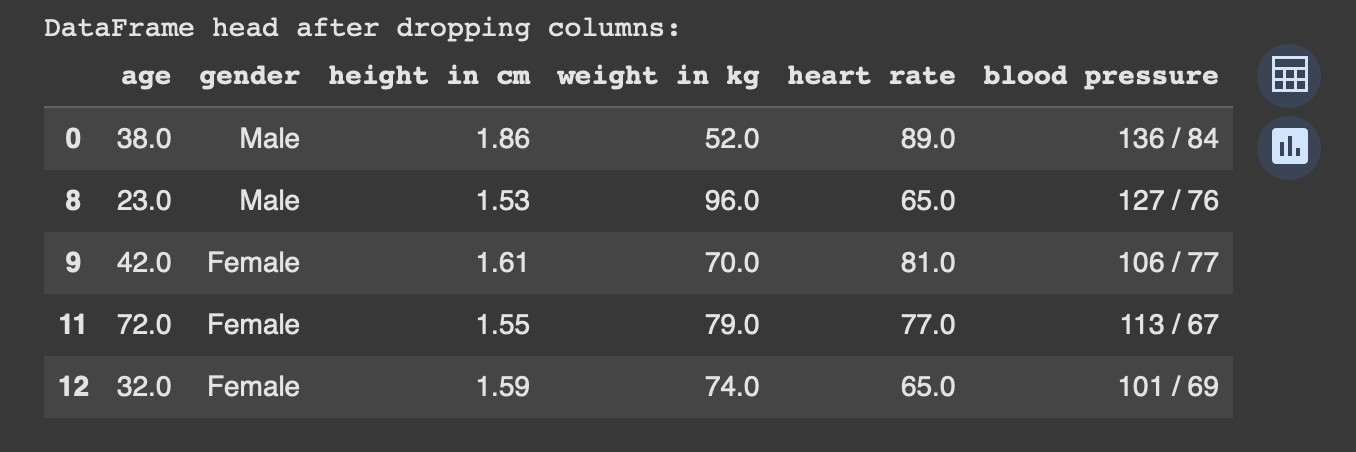
****

****

# Converting patient height to meters and standardise gender values :

****

****

****

**TASK 4 :**

Use AI to write a script to preprocess a social media text dataset.

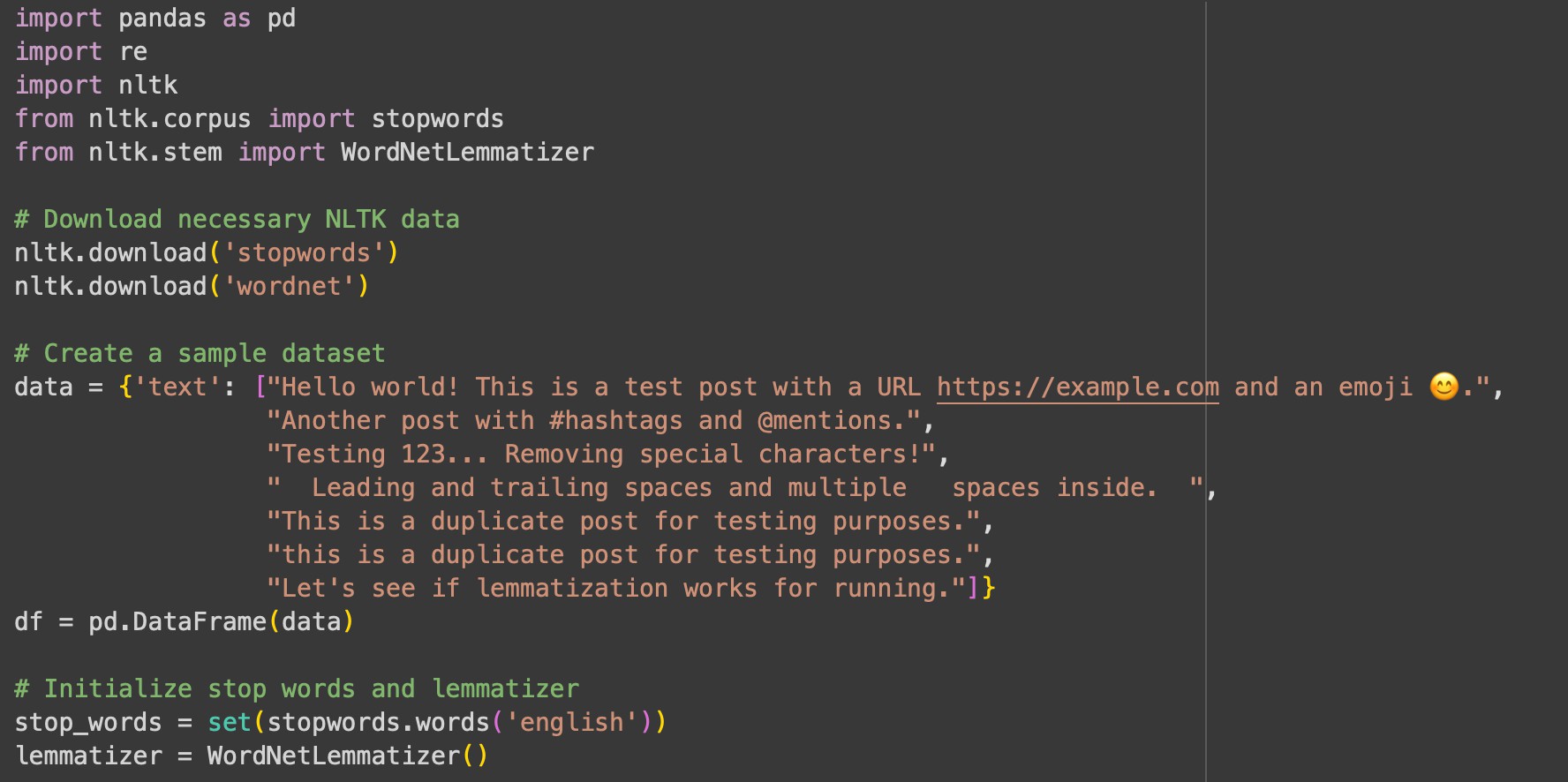
**Instructions:**

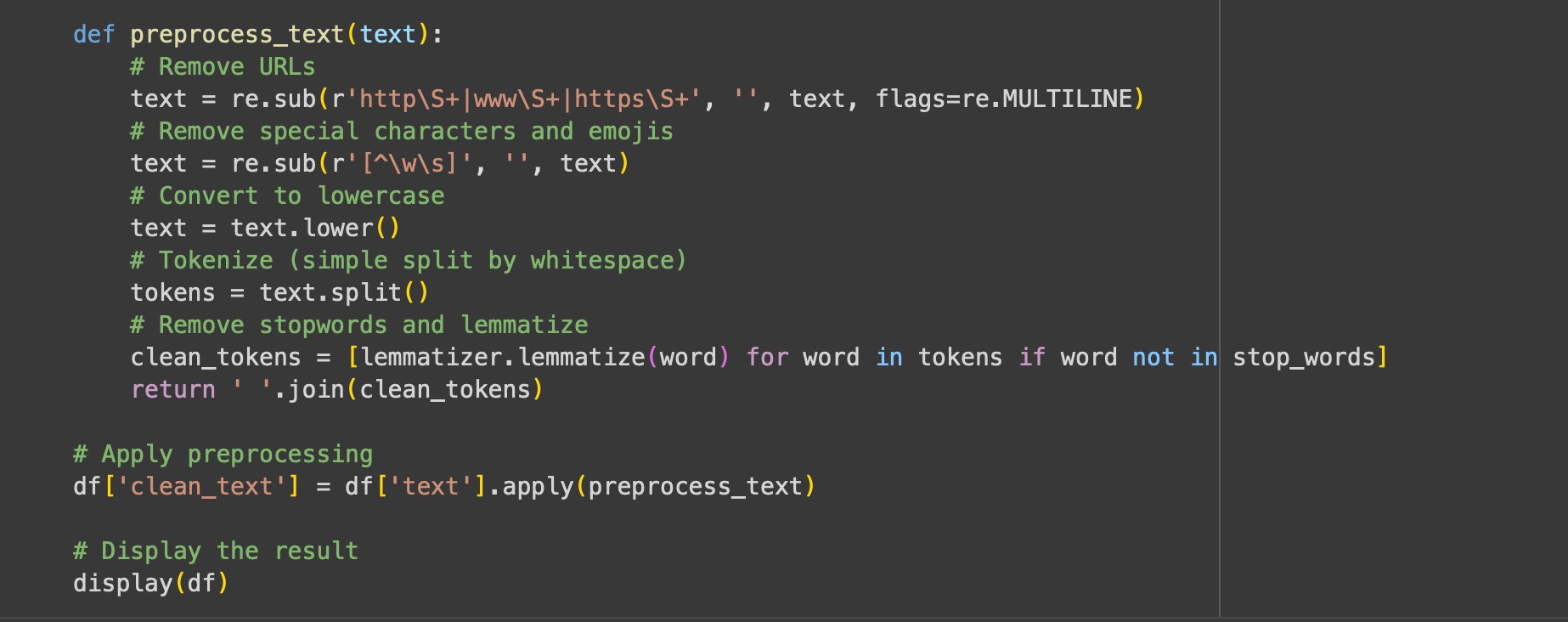
* Remove special characters, URLs, and emojis from text.
* Convert all text to lowercase.
* Tokenize and remove stopwords.
* Apply lemmatization for standardizing words.

**Expected Output:**

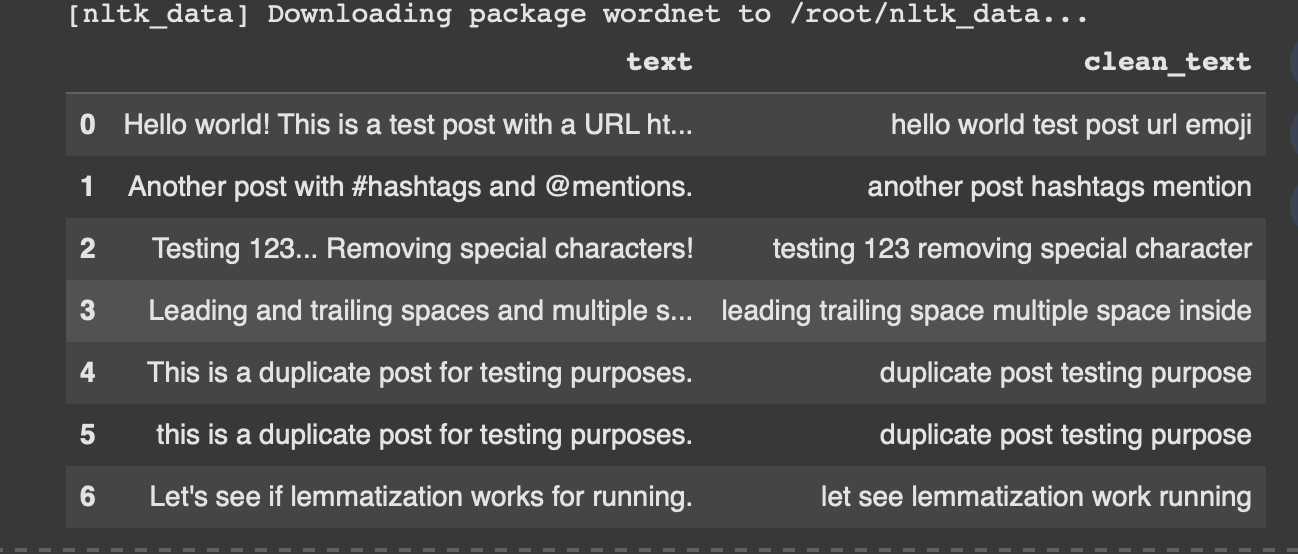
* A processed dataset with clean text, ready for NLP sentiment analysis.

**Expected code :**

****

****

# Preprocessed data :

****

**TASK 5 :**

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).

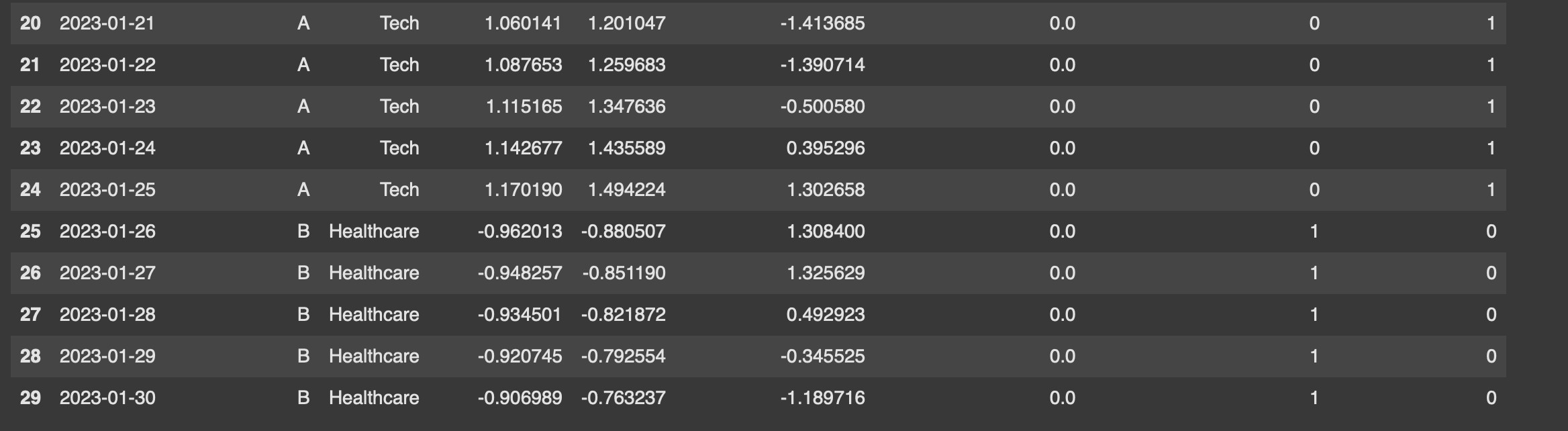
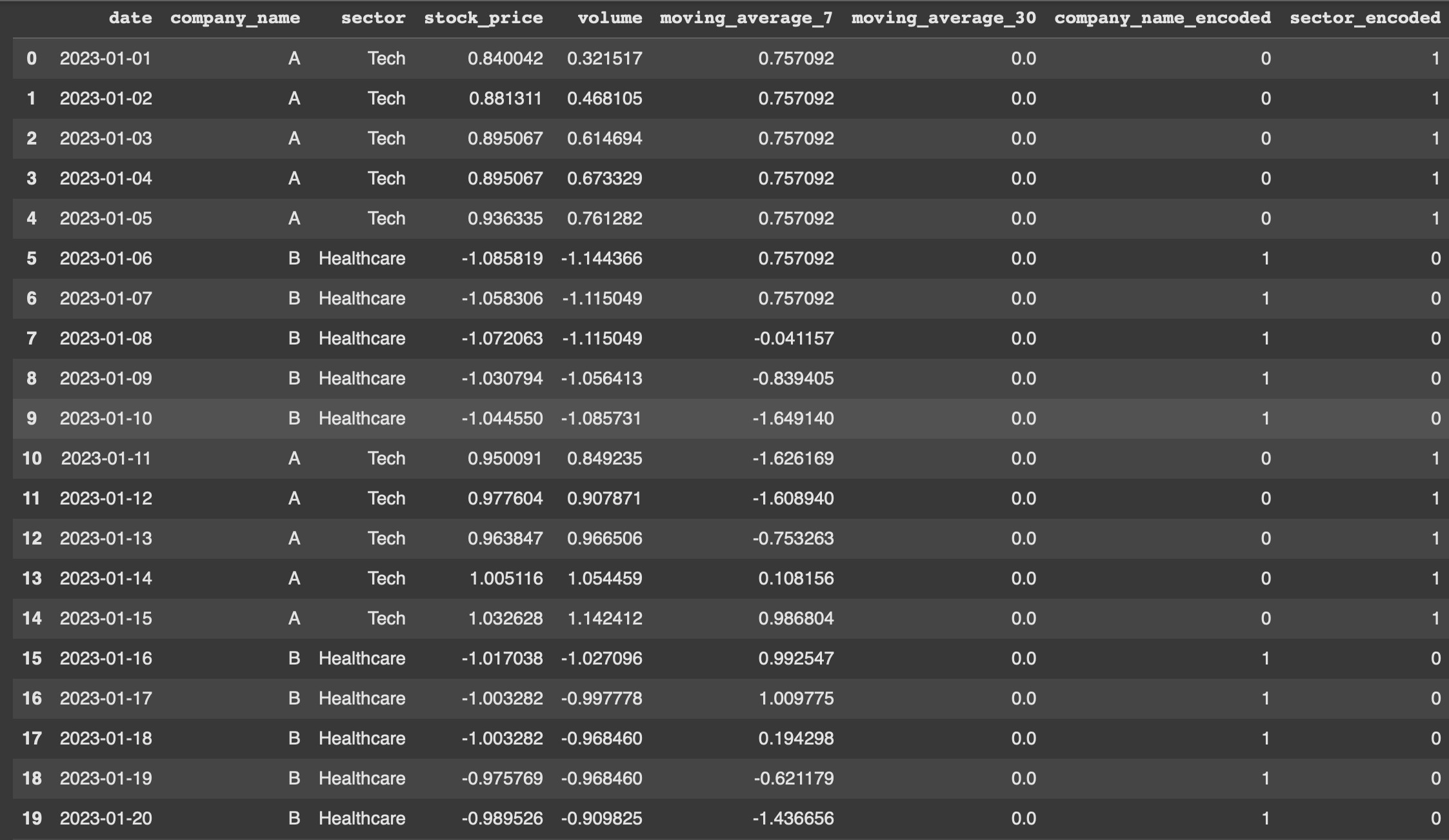
**Expected Output:**

* A feature-engineered DataFrame with new indicators and normalized values for ML tasks.





**Sample dataset with handling missing values and normalise continuous variables:**

****