**ASSIGNMENT 17**

**ENROLLNO:** 2503A51L33

**NAME:** T. SHIVA

**LAB:** AI ASSISTED CODING

**BATCH:** 20

**TASK 1**

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).

**PROMPT:**

Generate a Python script to clean an employee dataset with the following requirements:

* Handle missing values in the columns: salary, department, and joining\_date.
* Convert the joining\_date column to proper datetime format.
* Standardize department names (e.g., "HR", "hr", "Human Resources" → "HR").
* Encode categorical variables such as department and job\_role

**CODE:**

**A computer screen shot of a program

AI-generated content may be incorrect.**

**A computer screen with many colorful text

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**OUTPUT:**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**A black screen with white text

AI-generated content may be incorrect.**

**A screen shot of a computer

AI-generated content may be incorrect.**

**A screen shot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**OBSERVATION:**

** Missing Values Handled**:  
All missing values in salary, department, and joining\_date have been addressed. Salary was imputed using the median, department with the mode, and joining dates were filled with a placeholder and converted to datetime format.

* **Date Format Standardized**:  
  The joining\_date column is now in proper datetime format, enabling accurate time-based analysis.
* **Department Names Normalized**:  
  Variations like "hr", "HR", and "Human Resources" were standardized to "HR", ensuring consistency across records.
* **Categorical Encoding Applied**:  
  Categorical columns department and job\_role were label encoded, making them suitable for machine learning models.
* **Dataset Ready for Analysis**:  
  The cleaned dataset is now consistent, complete, and formatted for further exploration, visualization, or predictive modeling.

**TASK 2:**

Use AI to 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.

**PROMPT:**

Write a Python script to clean a sales dataset:

* Convert dates to datetime
* Add a 'Month-Year' column
* Remove rows with zero or negative amounts
* Normalize 'transaction\_amount' with Min-Max scaling

**CODE:**

A screenshot of a computer program

AI-generated content may be incorrect.

A screenshot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

**OUTPUT:**

A screenshot of a computer screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer screen

AI-generated content may be incorrect.

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screen shot of a computer

AI-generated content may be incorrect.**

**OBSERVATION:**

This preprocessing pipeline ensures clean, time-aware, and model-ready transaction data:

* Converting dates to datetime enables accurate time-based analysis.
* The 'Month-Year' column simplifies monthly aggregation and trend tracking.
* Removing non-positive amounts filters out invalid or refund-like entries.
* Min-Max scaling standardizes transaction values, making them suitable for machine learning models without distorting relative differences.

**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.

**PROMPT:**

Write a Python script to clean healthcare patient records:

* Fill missing values in numeric columns like blood\_pressure and heart\_rate using column mean.
* Convert height from centimeters to meters.
* Standardize gender labels (e.g., 'M', 'male', 'Male' → 'Male').
* Drop irrelevant columns like patient\_id after cleaning. Include comments for each step.

**CODE:**

**A screenshot of a computer program

AI-generated content may be incorrect.** A screenshot of a computer program

AI-generated content may be incorrect.**A screen shot of a computer program

AI-generated content may be incorrect.** **A screenshot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**OUTPUT:** **A screenshot of a computer screen

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.** **A screenshot of a computer screen

AI-generated content may be incorrect.**

**OBSERVATION:**

This cleaning process improves data quality and consistency for analysis:

* Filling missing numeric values with the mean preserves statistical integrity without introducing bias.
* Converting height to meters ensures unit standardization, crucial for modeling and comparisons.
* Standardizing gender labels eliminates redundancy and improves categorical clarity.
* Dropping irrelevant identifiers like patient\_id reduces noise and protects privacy in downstream tasks.

**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 stop words.  
• Apply lemmatization for standardizing words.

**PROMPT:**

Write a Python script to preprocess social media text:

* Remove special characters, URLs, and emojis
* Convert all text to lowercase
* Tokenize and remove stopwords
* Apply lemmatization to standardize words Include comments for each step.

**CODE:**



A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a program

AI-generated content may be incorrect.

**OUTPUT:**

**A computer screen shot of a black screen

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**OBSERVATION:**

This preprocessing pipeline transforms noisy, informal social media text into clean, structured input for NLP tasks:

* Removing special characters, URLs, and emojis reduces distraction and improves model focus.
* Lowercasing ensures uniformity across tokens, avoiding duplication due to case differences.
* Tokenization and stopword removal filter out common, non-informative words.
* Lemmatization standardizes word forms (e.g., “running” → “run”), enhancing semantic clarity and reducing vocabulary size.

**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).

**PROMPT:**

Write a Python script to preprocess a financial dataset:

* Fill missing values in stock price and volume
* Create 7-day and 30-day moving average features
* Normalize continuous variables using StandardScaler
* Encode categorical columns like sector and company\_name Include comments for each step.

**CODE:**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screenshot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**A screen shot of a computer program

AI-generated content may be incorrect.**

**OUTPUT:**

**A computer screen shot of a black screen

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screen shot of a computer

AI-generated content may be incorrect.**

**OBSERVATION:**

This preprocessing pipeline prepares financial data for robust analysis and modeling:

* Filling missing values in stock price and volume ensures continuity in time-series trends.
* Moving averages (7-day, 30-day) smooth out volatility and highlight underlying patterns.
* StandardScaler normalizes continuous features, improving model performance and convergence.
* Encoding categorical variables like sector and company\_name transforms qualitative data into numeric form, enabling compatibility with machine learning algorithms.