ASSIGNMENT-17

NAME:K.RITHIKA

ROLL.NO:2403A510G9

TASK1:

PROMPT:

“Clean the dataset of social media posts by doing the following:

1. Remove stop-words, punctuation and special symbols from the post-text.
2. Handle missing values in the likes and shares columns (e.g., impute or drop).
3. Convert the timestamp column into a datetime type, then extract the hour and weekday features.
4. Detect and remove spam and duplicate posts.  
   Output the cleaned dataset ready for sentiment and engagement analysis.”

CODE:

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a program code

AI-generated content may be incorrect.

A computer screen with text

AI-generated content may be incorrect.

OUTPUT:

A computer screen shot of a computer program

AI-generated content may be incorrect.

OBSERVATIONS:

.Number of rows before cleaning vs after duplicate/spam removal.

.Number of missing values in likes, shares before imputation and how many were filled.

.Before/after average text length (in words) to see effect of cleaning.

.Engagement by hour/weekday (e.g., posts at 14 h get higher average likes).

.Distribution of cleaned text: how many posts now have zero words (i.e., originally blank or after cleaning became empty) → you may want to drop those.

.Any obvious spam user patterns: e.g., a small number of users contributing a large fraction of posts or many posts with identical cleaned text.

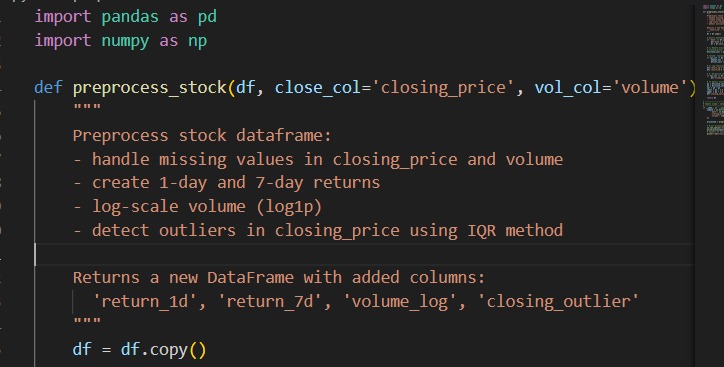
TASK-2:

PROMPT:

WRITE THE CODE BY FOLLOWING THE INSTRUCTIONS:

Handle missing values in closing\_price and volume.  
- Create lag features (1-day, 7-day returns).  
- Normalize volume column using log-scaling.  
- Detect outliers in closing\_price using IQR Model

CODE:



A computer screen with text on it

AI-generated content may be incorrect.

A screen shot of a computer screen

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.

OBSERVATIONS:

.How many missing values there were in closing\_price and volume, and how many we filled.

.What percentage of the data was used for modelling after creating lags (i.e., initial rows lost).

.Before and after log-scaling: e.g., average and median of volume, how skew it was, and how it changed after log.

.How many outlier days we flagged using IQR for closing\_price, and whether they correspond to major market events or data errors.

.Whether the returns (1-day and 7-day) look reasonable: mean, standard deviation, maybe extreme values.

. A check: do large volumes correspond to large returns (either positive or negative)? That might show meaningful patterns.

TASK-3:

PROMT:

WRITE THE CODE BY FOLLOWING THE INSTRUCTIONS:

-Handle missing values using forward fill.  
- Remove sensor drift (apply rolling mean).  
- Normalize readings using standard scaling.   
- Encode categorical sensor IDs.

CODE:

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen shot of a program code

AI-generated content may be incorrect.

A screen shot of a computer program

AI-generated content may be incorrect.

A computer screen with colorful text

AI-generated content may be incorrect.

A screen shot of a computer code

AI-generated content may be incorrect.

OUTPUT:

A screenshot of a computer program

AI-generated content may be incorrect.

A black screen with white text

AI-generated content may be incorrect.

OBSERVATIONS:

.How many missing readings we had per sensor, and how many were filled by forward-fill.

.After rolling mean, how much the variance of each sensor’s readings dropped (i.e., drift decreased).

.Before and after scaling: what were the raw means/variances of temperature/humidity vs the scaled values.

.How many sensor IDs we have, and how evenly the readings are distributed across sensors after encoding.

.Any sensors whose readings still deviate strongly from the scaled mean (possible faulty sensors or outliers).

TASK-4:

PROMPT:

WRITE THE CODE BY FOLLOWING THE INSTRUCTIONS:

Standardize text (lowercase, remove HTML tags).  
- Tokenize and encode reviews using AI-assisted methods (TF-IDF or  
embeddings).

- Handle missing ratings (fill with median).  
- Normalize ratings (0–10 → 0–1 scale).  
- Generate a before vs after summary report

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 computer screen shot of a program

AI-generated content may be incorrect.

OUTPUT:

A black screen with white text

AI-generated content may be incorrect.

A computer screen with white text

AI-generated content may be incorrect.

OBSERVATIONS:

.We made all review text lowercase and removed HTML tags so the text is consistent and clean.

.We tokenized and encoded the review text (using TF-IDF or embeddings) so the model can understand the meaning behind the words.

.We filled missing ratings with the median and scaled ratings from 0-10 down to 0-1 so every review has a usable, normalized score.

.We compared “before vs after” cleaning to see how many reviews were missing data, .how messy the text looked before, and to confirm the dataset is ready for a sentiment-classification model.