Solution Plan for Data Preprocessing

1. Data Loading and Initial Exploration

- **Loaded dataset** using `pandas`
- **Checked dataset structure** using:
  - `data.shape` → Dimensions of the dataset.
  - `data.info()` → Data types and non-null counts.
  - `data.head()` → First few rows of the dataset.

---

2. Data Cleaning Steps

Step 1: Handling Missing Values

- Checked missing values using `data.isnull().sum()`.
- Dropped rows where `"Scrape Timestamp"` had invalid dates.
- Converted `"Scrape Timestamp"` to datetime using:
  python
- Removed rows with missing timestamps using:
  python

Step 2: Handling Duplicates

- Checked duplicate rows.
- No explicit duplicate removal was performed.

Step 3: Feature Engineering

- Extracted `"Date"` from `"Scrape Timestamp"`.

---

3. Data Visualization & Exploration

- **Rating Distribution:** Used `sns.countplot()` to analyze rating frequencies.
- **Platform Analysis:** Used `sns.barplot()` to analyze review sources.
- **Word Frequency Analysis:**
  - Tokenized and counted words from `"Review Paragraph"`.
  - Visualized most common words using a bar chart.
- **Word Cloud Generation:** Created a word cloud from review texts

4. Text Preprocessing (for NLP Models)

- Tokenized review texts.
- Removed common stopwords (not explicitly mentioned but likely needed).
- Generated word clouds and frequency distributions.

Further Steps:

NLP Sentiment Scoring Methods

Lexicon-Based Sentiment Analysis:

- NLTK Opinion Lexicon
- VADER Sentiment Scoring

Transformer-Based Sentiment Analysis

- Sentiment Scoring using RoBERTa
- Sentiment Scoring using Transformers Pipeline

Evaluation Metrics

- **Confusion Matrix (`confusion_matrix()`.** Compares predicted vs. actual sentiment labels.
- **Accuracy** – % of correctly classified reviews.
- **Precision & Recall** – Important when class imbalance exists (e.g., more positive reviews than negative).
- **F1-Score** – Balance between precision and recall.
- **AUC-ROC**

SS