Data Cleaning Questions

Based on Mobile Reviews Sentiment.csv

General Structure

1. Check for and remove duplicate rows (e.g., same `review_id` or repeated reviews).

2. Verify that each `review_id` is unique.

3. Drop unnecessary columns if they are not useful for analysis.

Text Cleaning

1. Standardize the `customer_name` field (remove leading/trailing spaces, consistent casing).

2. Normalize the `sentiment` column to consistent categories (e.g., `positive`, `negative`, `neutral`).

3. Ensure the `language` column values are valid ISO codes (e.g., `en`, `es`).

Date & Time

1. Convert `review_date` into a proper date format (`YYYY-MM-DD`).

2. Extract **year**, **month**, and **day** into separate columns.

3. Check for missing or inconsistent dates (e.g., future dates).

Numerical Consistency

1. Ensure `price_usd`, `price_local`, and `exchange_rate_to_usd` align correctly (`price_local / exchange_rate_to_usd = price_usd`).

2. Check that all rating columns (`battery_life_rating`, `camera_rating`, `performance_rating`, `design_rating`, `display_rating`) are within a valid range (e.g., 1–5).

3. Validate that `review_length` and `word_count` match the content of `review_text`.

4. Ensure `helpful_votes` are non-negative integers.

Categorical Data

1. Standardize `brand` and `model` names (e.g., unify `Samsung` vs `SAMSUNG`).
2. Verify that all `currency` values are valid (e.g., `USD`, `INR`, `BRL`, etc.).
3. Clean the `source` column (standardize platform names like `Amazon`, `Flipkart`, etc.).

New Features

1. Create a binary column for `verified_purchase` (1 = True, 0 = False).
2. Derive a **review sentiment score** from the categorical sentiment labels.
3. Create an overall **average rating** per review from all rating sub-categories.
4. Add a column categorizing reviews into `short`, `medium`, or `long` based on `word_count`.