Pharmaceutical Products Dataset

Data Cleaning, Correction, and Transformation Practice Questions

■ 1. Identify and remove duplicate rows based on Product_ID.

■ 2. Check for missing values in 'Expiry_Date' and replace them with a placeholder date.

■ 3. Standardize the 'Supplier' names (e.g., 'Cipla ' with trailing spaces should be corrected to 'Cipla').

■ 4. Convert 'Order_Date' and 'Expiry_Date' into proper Date format.

■ 5. Create a new column to calculate 'Total_Sales' = Unit_Price × Quantity.

■ 6. Split the 'Strength' column into two: 'Strength_Value' and 'Strength_Unit'.

■ 7. Replace outlier values in 'Unit_Price' (greater than 400) with the average price of that category.

■ 8. Correct any inconsistent 'Form' entries (e.g., 'tablet', 'Tablet ', 'TAB' → 'Tablet').

■ 9. Identify expired products using 'Expiry_Date' and mark them as 'Expired' or 'Valid'.

■ 10. Group data by 'Category' and calculate average 'Unit_Price'.

■ 11. Detect and correct inconsistent country names (e.g., 'U.S.A', 'USA ').

■ 12. Normalize 'Popularity_%' to ensure all values lie between 0 and 100.

■ 13. Create a calculated column for 'Remaining_Shelf_Life' in days = Expiry_Date - Today.

■ 14. Transform the dataset to show Yearly Sales by extracting year from 'Order_Date'.

■ 15. Create a column that categorizes products as 'High Demand' if Popularity_% > 70, else 'Low Demand'.

■ 16. Check for negative or zero 'Quantity' values and correct them.

■ 17. Merge supplier details (create a lookup table for Supplier with Supplier_Country).

■ 18. Pivot the dataset to show 'Form' as columns and 'Quantity' as values.

■ 19. Unpivot the pivoted data back to long format for Power BI modeling.

■ 20. Create a calculated column 'Revenue_Category' that buckets Total_Sales into Low (<5000), Medium (5000-20000), High (>20000).