Feature Engineering

In this task, you will be performing feature engineering by adding new columns to an existing dataset. You are required to create four new columns by following the instructions below.

Step 1: Categorize Age into Age Groups

Objective: Create a new column `age_group` by categorizing the existing `age` column into specific age ranges.

- Categorize ages into the following groups:
  - 18-25
  - 26-35
  - 36-45
  - 46-55
  - 56-70
  - 70+
- Ensure that each age in the dataset is mapped to one of these age groups.
- After that, remove the original `age` column from the dataset.

Step 2: Create `cf_ab_score` (Consume Frequency and Awareness Brand Score)

Objective: Create a new column `cf_ab_score` by combining the information from `consume_frequency(weekly)` and `awareness_of_other_brands` columns.

- Use the following mappings for `consume_frequency(weekly)`:
  - "0-2 times" → 1
  - "3-4 times" → 2
  - "5-7 times" → 3
- Use the following mappings for `awareness_of_other_brands`:
  - "0 to 1" → 1
  - "2 to 4" → 2

    - "above 4" → 3
- Calculate `cf_ab_score` using the following formula:

$$\text{cf\_ab\_score} = \frac{\text{frequency score}}{\text{awareness score} + \text{frequency score}}$$

- Round the result to two decimal places.

Step 3: Create Zone Affluence Score (ZAS)

Objective: Calculate the `zas_score` using a combination of the `zone` and `income_levels` columns.

- Use the following mappings for the `zone` column:
    - "Urban" → 3
    - "Metro" → 4
    - "Rural" → 1
    - "Semi-Urban" → 2
- Use the following mappings for the `income_levels` column:
    - "<10L" → 1
    - "10L - 15L" → 2
    - "16L - 25L" → 3
    - "26L - 35L" → 4
    - "> 35L" → 5
    - "Not Reported" → 0
- Calculate the `zas_score` using the following formula:

$$\text{zas\_score} = \text{zone score} \times \text{income score}$$

Step 4: Brand Switching Indicator (BSI)

Objective: Create a binary indicator column `bsi` that identifies if a respondent is likely to switch brands.

- Check if the respondent's `current_brand` is not "Established".
- Also check if the `reasons_for_choosing_brands` are either "Price" or "Quality".
- If both conditions are true, assign a value of `1` to indicate potential for brand switching. Otherwise, assign `0`.

Final Cleaning Step:

Removing Logical Outliers:

- When reviewing the occupation data, we found logical inconsistencies. For instance, there are students listed in the `56-70` age group, which seems like an incorrect entry. We need to remove such records where the data doesn't make sense logically.

occupation	Entrepreneur	Retired	Student	Working Professional
age_group				
18-25	535	0	7328	2605
26-35	1826	0	697	6570
36-45	1619	0	0	4353
46-55	799	0	0	2167
56-70	221	1130	35	106
