Data Cleaning Instructions

In this phase, you will clean the survey data to ensure it's ready for modeling. Follow the guidelines below to help you structure your approach. While some hints are provided, you are encouraged to explore the best methods to handle these tasks.

Step1: Remove Duplicates:

- Investigate if the dataset contains duplicate entries. Consider which columns could help identify duplicates. Think about how duplicate records might affect the outcome of the model and remove them accordingly.

Step2: Outlier Detection in Age:

- Explore the \`age\` column to spot potential outliers. You can use statistical methods or visualizations (e.g., box plots) to help identify these. Reflect on why it's important to handle these outliers. Based on your findings, decide whether to keep, adjust, or remove them.

Step3: Handling Missing Data:

- For the \`income_levels\` column, missing values can be problematic. What would be a reasonable way to replace them? (Hint: Consider using "Not Reported" for missing income levels)
- For the \`consume_frequency(weekly)\` and \`purchase_channel\` columns, think about the best way to fill missing values. Could the most common (mode) values be an appropriate replacement? Investigate the distributions before making a decision.

Step4: Correcting Spelling Mistakes in Categorical Data:

- Review the entries in the \`zone\` and \`current_brand\` columns. Are there any inconsistencies in spelling or formatting? How could you identify and fix these issues to ensure uniform categories?