Hello,

I've reviewed the quality of our Users, Products, and Transactions datasets and identified several key issues. Below are the main findings, along with some recommended solutions for certain issues. Once we align on the solutions, I will continue with the data cleaning process. I've also highlighted areas where I need your input and assistance.

Common Issues:

- Missing Data:

  Critical fields like **BARCODE** (product code), **BIRTH_DATE**, and **FINAL_SALE** have significant missing data, which may affect data merging and analysis. I can clean these missing values once approved.

- Duplicated Rows:

  - Products: ~400 duplicated rows. I recommend removing these to prevent issues during data merging.
  - Transactions: ~320 duplicated rows. It's unclear whether these are errors or if the same transaction can appear multiple times in a receipt. Your input here would be helpful.

- Data Integrity:

  After merging the datasets, only ~144 rows remain, significantly down from the original 50,000 in Transactions, 100,000 in Users, and 845,000 in Products. The main issue appears to be that fewer than 100 User IDs in the Users dataset match the Transaction dataset, excluding much of the data. This will likely cause problems if we use such a small proportion to represent the entire trend. I need your assistance to review the data integrity or let me know if further manipulation is required.

Dataset-Specific Issues:

Users Data:

- The most common **BIRTH_DATE** value is 1970-01-01 (Unix epoch date), likely placeholders. I suggest replacing these with NaN.
- Unusually low or high ages (e.g., below 13 or above 100). I suggest applying age limits with a minimum of 13 and maximum of 100 for data accuracy.

Products Data:

- There are 27 duplicated **BARCODE** entries with different values in other attributes like brand and manufacturer. I need your input on how to handle these discrepancies.

Transactions Data:

1. Approximately 11,000 transactions have the same **RECEIPT_ID** and **BARCODE** but different **FINAL_QUANTITY** or **FINAL_SALE** values. I am unclear about the cause of these discrepancies. Why would the same product appear multiple times on the same receipt with different values? Your input here would be helpful.
2. There are 0 and non-whole numbers (e.g., 6.22, 1.23) in **FINAL_QUANTITY** in transactions. It would be helpful if you can explain the cause of these issues.

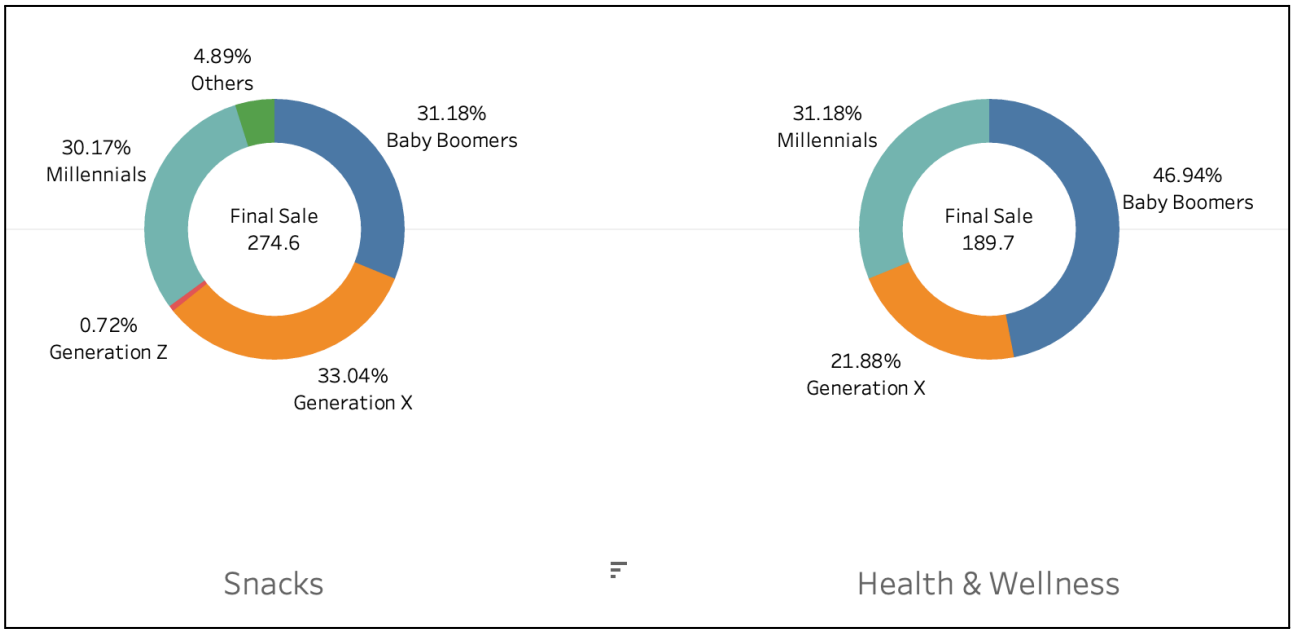
3. There are 0 and missing values in **FINAL_SALE** in transactions. It would be helpful if you can clarify the cause of this situation.

Interesting Trend:

While investigating the data, I also found some interesting trends that I'd like to share with you. Please note that, as mentioned earlier, after merging the datasets, only 144 rows remain, which is a small proportion of our data. Additionally, the transaction dataset covers only the period from June to September 2024, so these findings are based on a three-month span and a limited data sample. I'll be happy to provide more insights once we've completed the data cleaning and have a more robust dataset.

From the data, I observed that in the top two categories with the highest sales—**Snacks** and **Health & Wellness**—the dominant consumers are **Baby Boomers, Generation X, and Millennials**. This insight suggests that the customer base for these product categories is primarily made up of a more mature audience.

I've attached a pie chart below for your reference:



Next Steps:

I'd really appreciate your input with the issues I've highlighted in yellow above. For the other data quality issues, I've put together a plan to address them, and I'd love to schedule a meeting with you to make sure my approach makes sense and to answer any questions you might have. Your input would be incredibly helpful, and I'm excited to hear your thoughts!

Looking forward to hearing from you and finding a time to chat.

Best regards,
Mandy Liou