Subject: Data Quality Insights and Key Findings from Our Investigation

Hi [Recipient's Name],

I wanted to share a quick summary of our recent data investigation, highlighting key issues, an interesting trend, and a request for further support to improve our analysis.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Key Data Quality Issues:

1. Missing and Inconsistent Data:
   * Over 43% of products in the transaction data cannot be directly matched with the product table, making it challenging to accurately analyze product sales.
   * For user data, 99% of user IDs in the transaction data cannot be matched with the user dataset, potentially affecting user behavior analysis.
   * Inconsistencies in Quantity and Price in the transaction dataset could lead to inaccuracies in calculating total sales and profits.
2. Ambiguous Data Definitions:
   * Columns like CATEGORY\_1 through CATEGORY\_4 in the product dataset lack clear definitions. Based on the data, they appear to represent a hierarchy (e.g., department, category, subcategory, and product type), but this requires confirmation.
   * Date-related fields (e.g., transaction dates) are inconsistently formatted, complicating future analysis and requiring standardization.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Interesting Findings related to customer generations:

1. Millennials Drive the Majority of Sales:

* Millennials account for 49.53% of total sales, making them the largest contributor among all generations. Baby Boomers follow at 29.87%, and Gen X contributes 20.6%.

1. Spending Behavior Differences:

* Millennials not only dominate sales but also show a higher average spend per person ($12.41) and average receipts (3.31) compared to other generations. However, their spending per receipt ($3.75) is slightly lower than Baby Boomers.
* Baby Boomers have the highest spending per receipt ($4.02), indicating that while they shop less frequently (2.86 average receipts), their transactions tend to have higher value.
* Gen Z exhibits relatively high shopping frequency (4.00 average receipts) but lower spending per receipt ($2.74), suggesting smaller basket sizes.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Request for Action:

To make sense of the data and address outstanding issues, we need the following:

1. Clarification on Data Definitions:
   * Confirm the intended structure and use of CATEGORY\_1 through CATEGORY\_4 in Product dataset.
   * Clarify the significance of unmatched user IDs and whether they are expected (e.g., new users not yet synced).
2. Support for Data Cleanup:
   * Address inconsistencies in FINAL\_QUANTITY and FINAL\_SALE in Transaction dataset.
   * Ensure product and user tables are reconciled with transaction data.
3. Additional Context:
   * What specific metrics or insights are most valuable to the business? This will help us prioritize cleanup and further analysis.

Let me know if you’d like to discuss further or if there’s a good time to align on next steps.

Best regards,  
Chunwei