Subject: Data Quality Findings and Next Steps

Hi Business Leader,

I hope you're having a wonderful day! I wanted to share some key findings from our recent data investigation, along with some outstanding questions and recommendations.

Key Data Quality Issues:

1. **Missing Data:**
   - The **Products dataset** has missing values in key fields like CATEGORY_4, MANUFACTURER, and BRAND, which may impact product-level analysis.
   - The **Transactions dataset** has significant missing values in BARCODE and FINAL_SALE. The high number of missing FINAL_SALE values could limit our ability to analyze revenue trends.
   - The **USERS dataset** has missing values in GENDER, LANGUAGE, and STATE, with LANGUAGE having a particularly high percentage of missing data. This may affect demographic insights.
2. **Data Consistency & Cleaning Needs:**
   - The FINAL_QUANTITY field includes non-numeric values like "zero," which need conversion for accurate calculations.
   - We identified **duplicate records**, which may require de-duplication to ensure accurate reporting.
   - **AGE distribution anomalies** suggest some users have unrealistic birth years (100+ years old), likely due to incorrect or missing data.

Interesting Trend in the Data:

Despite the data quality challenges, one notable finding is that **Millennials (1981-1996) account for the highest percentage of purchases in the Health & Wellness category**. This suggests that younger generations are more engaged with wellness products, which could inform targeted marketing strategies.

Request for Action:

To improve data reliability and extract more meaningful insights, we need:

1. **Clarification on the missing FINAL_SALE values**—Are these truly missing transactions, or should we assume zero sales for those cases?
2. **Guidance on handling extreme AGE values—**Should we apply a reasonable cutoff (e.g., users over 100 years old as outliers)?
3. **Confirmation on how duplicate records should be resolved**—Should we de-duplicate based on USER_ID, BARCODE, or another unique identifier?
4. **Support from the data engineering team** to clean and standardize FINAL_QUANTITY, ensuring all values are numeric.

Would love to discuss this further and align on next steps. Let me know how you'd like to proceed.

Many Thanks,
Sherry