Fourth Question:

Subject: Data Quality Assessment Findings & Next Steps

Respected Ma’am/Sir,

I hope you are doing well. I wanted to share key findings from our recent data quality assessment across the **PRODUCT_TAKEHOME, TRANSACTION_TAKEHOME, and USER_TAKEHOME** datasets. Below are some key insights, trends, and outstanding questions that may require further discussion.

Our analysis revealed several data quality issues that may impact the accuracy of reporting and decision-making.

In the **PRODUCT_TAKEHOME dataset**, missing descriptions and additional product details limit the depth of analysis. CATEGORY_4 has the highest number of missing values, but if this is an optional field, its impact on data integrity may be minimal. However, CATEGORY_1 contains null values, making it difficult to classify products within transactions, which could affect future analysis.

The **TRANSACTION_TAKEHOME dataset** contains a significant number of void transactions, with only 25% of the data being usable. Additionally, missing values in FINAL_QUANTITY, FINAL_SALE, and BARCODE introduce inconsistencies in transaction tracking.

In the **USER_TAKEHOME dataset**, out of 100,000 users, only 262 transactions could be traced, and only 72 transactions were valid. Furthermore, only 59 users have valid transactions, raising concerns about data completeness and engagement levels.

While reviewing the dataset, we identified some interesting trends. The top five states with the most active users are **Texas, Florida, California, New York, and Illinois**, highlighting potential geographic hotspots for user engagement. Additionally, the highest sales are recorded at **Walmart, Costco, Sam’s Club, CVS, and Target**, which may provide insights into consumer shopping patterns and store performance.

To improve data quality and usability, we recommend enhancing barcode scanning validation to ensure that complete receipt details are captured. Addressing CATEGORY_1 null values by exploring whether they can be inferred from other product attributes may also improve data completeness. Given the volume of void transactions, it is important to determine if they should be excluded from key analysis or flagged for a specific business purpose. Additionally, implementing a reward system for users who scan receipts could increase engagement and improve data tracking. To further standardize the dataset, CATEGORY_1 values should be normalized, and store names should be cleaned to remove inconsistencies (e.g., "Wal-Mart" vs. "Walmart"). These improvements will ensure more accurate reporting and enhance data-driven decisions.

To move forward, we need clarification on a few key areas. What does CATEGORY_1 represent, and should missing values be addressed? If not, can the product be inferred from other categories? How are BARCODE values captured, and how should we handle duplicate barcodes associated with different brands and manufacturers? Lastly, do void transactions follow a specific pattern, or should they be excluded entirely from the analysis?

I would love to hear your thoughts on these points. Happy to discuss this further!

Best,
Shruti