Fourth Question:

Subject: Data Quality Assessment Findings & Next Steps

Respected Ma'am/Sir,

I wanted to share key findings from our recent data quality review across the PRODUCT_TAKEHOME, TRANSACTION_TAKEHOME, and USER_TAKEHOME datasets. Below are some insights and open questions that may require further discussion.

Key Data Quality Issues:

PRODUCT_TAKEHOME Table:

- Some descriptions and additional details are missing, limiting our ability to conduct a deeper analysis.
- CATEGORY_4 has the most missing values, but if it's an optional field, it may not impact data integrity.
- CATEGORY_1 contains null values, making it difficult to accurately identify products within transactions.

TRANSACTION_TAKEHOME Table:

- Void transactions make up a significant portion of the dataset, with only 25% of data being usable.
- Missing values in FINAL_QUANTITY, FINAL_SALE, and BARCODE introduce inconsistencies in transaction tracking.

USER_TAKEHOME Table:

- Out of 100,000 users, only 262 transactions can be traced, and only 72 transactions are valid.
- Only 59 users have valid transactions, raising questions about data completeness and engagement levels.

Interesting Trend Identified:

- Geographic Hotspots: The top five states with the most active users are Texas, Florida, California, New York, and Illinois.
- Top Performing Stores: The highest sales are recorded at Walmart, Costco, Sam's Club, CVS, and Target.

Recommendation:

Improve Data Collection & Accuracy

- Enhance barcode scanning validation to ensure complete receipt details are captured.
- Address CATEGORY_1 null values by investigating if they can be inferred from other product attributes.

Handle Void Transactions More Effectively

- If void transactions contain incomplete or incorrect data, they should be flagged or removed from key analyses.

Increase User Engagement & Data Coverage

- Implement a reward system where users earn incentives for scanning receipts, improving data completeness.
- Provide real-time feedback after a scan to prompt users to capture missing details.

Improve Data Cleaning & Standardization

- Standardize CATEGORY_1 values to ensure proper product classification.
- Normalize store names (e.g., "Wal-Mart" vs. "Walmart") to remove inconsistencies.

Leverage Insights for Business Strategy

- The top-performing states and stores could be used to optimize promotions, discounts, or regional campaigns.
- If certain stores have high transactions but low user count, investigate whether this indicates bulk purchases or specific shopping behaviors.

Request for Action:

To move forward, we need more clarification on the following:

- What does CATEGORY_1 represent, and should missing values be addressed? If not, can the product be inferred from other categories?
- Can we get more details on how BARCODE values are captured? How do we tackle the duplicate barcodes that have different brands and manufacturers?
- Do void transactions follow a specific pattern, or should they be excluded entirely from analysis?

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

Best,
Shruti