Communicate with stakeholders

Subject: Action Needed: Data Quality Issues and Key Growth Trend Insights

Hello [Leader's Name],

In my review of our rewards app datasets, I identified data quality and consistency issues that could impact the value of our insights. Addressing these will strengthen the accuracy of actionable findings.

Here’s a summary of key points and clarifying questions.

Data Quality Issues:

- **Missing Data:** High levels of missing data were noted in several fields: language(30.5%), gender(6%), birth date (4%) in Users , category_4 (92%), manufacturer/brand (26.8%) in Products, and barcode (11.5%) and sale (25%) in Transactions.
- **Validity Concerns:** Approximately 18% of user IDs in the Transactions dataset are not present in the Users dataset, and 4% of barcodes in Transactions are missing from Products. These are critical field for linking products can lead to to inaccuracies in understanding popular brands, categories, and total revenue distribution.
- **Duplicates:** There are 0.5% duplicates in Product barcodes and 51.2% in Transaction receipt IDs, potentially causing inaccuracies in understanding popular brands, categories, and total revenue distribution.

Data Inconsistencies:

- **Users:** Several birth dates are invalid, with outliers and a large number of entries showing a default date of "1970-01-01."
- **Products:** Barcode lengths vary widely (5-16 digits), and there are formatting inconsistencies in the category fields.

- **Transactions:** We found 94 records where the scan date occurs before the purchase date, and ~25% (12,500) records with a quantity labeled as "zero."

Outstanding Questions:

1. **Users Dataset:** Could we get clarity on the user_id criteria, including which user types are included or excluded?
2. **Products Dataset:** Are barcodes meant to be unique? Additionally, we need a better understanding of the category structure and the term “Placeholder Manufacturer” in manufacturer
3. **Transactions Dataset:** Could you confirm if receipt_ids are intended to be unique? Also, insights into "zero" quantities with non-zero sales and instances of purchase dates after scan dates would be helpful.

These steps will help us close data gaps and provide stronger, more actionable insights. Please let me know if there’s someone I can connect with for further clarification or documentation.

Thanks,

Ramya Kalyanaraman

Insight Architect