Subject: Data Quality Issues Identified in Receipt Data - Next Steps & Questions

Hi Team,

I wanted to bring some important data quality issues to your attention that I've encountered while working with the data. As these issues might impact the reporting, customer experience, and overall decision-making, I think it's essential we address them. Below is an overview of the findings, as well as some questions to guide next steps.

Summary of issues:

Key data quality issues include missing critical fields across multiple tables, orphaned records with unmatched barcodes, and a high percentage (55%) of missing barcodes in the ReceiptItems table, which could affect product identification, customer activity analysis, and reporting accuracy.

How Did I Discover These Issues?

I ran a series of SQL queries and python scripts using pandas across the dataset to check for missing values, orphaned records, and inconsistent data types. This revealed the gaps and discrepancies that could affect the quality of our analyses and business insights.

Key Questions:

Data Sources: Where is the most up-to-date brand and product information coming from? Are there any recent changes or gaps in how this data is collected or maintained?

Missing Fields: For the fields with missing values , do you have any insight into why they might be empty or misreported? Are there any known processes or external systems that might affect these fields?

Orphaned Records: We'll need to understand why some receipt items do not match with a brand in our records. This might involve cleaning up barcode data or correcting mapping errors.

Missing Barcodes: We need to investigate the root cause of these missing values and whether this is a data entry issue or a product catalog issue. Are we using a standardized barcode system across all products, or is there any possibility of mismatched or outdated barcodes in our data?

Next Steps:

Data Cleanup: We'll need to fill in missing data, especially for fields crucial for product and customer analysis

Orphan Record Investigation: Work on understanding why some receipt items don't have a corresponding brand and help resolve these mismatches.

Improvement of Barcode Collection: Addressing the missing barcodes will likely involve improving the data collection process or exploring alternative methods for identifying products.

Information Needed to Resolve Issues:

Access to the most recent dataset for both receipts and brand information to cross-check. Insights into any external factors or systems affecting barcode and brand data synchronization.

Guidance on any business rules that could clarify whether missing brand data or incorrect fields should be flagged or excluded from analysis.

Performance & Scaling Considerations:

As our dataset grows, we might face performance issues, especially when performing joins and aggregations across large tables. To address this, we can optimize queries, ensure proper indexing and monitor performance regularly.

It would also be helpful to set up automated data validation checks as new data enters the system, helping to catch issues early on.

Please let me know if you have any questions or need further clarification.

Best regards,

Ramya Ravikanti.