

Subject: Data Quality Issues and Insights from Recent Investigation⁽³⁾

Hi [Stakeholder's Name],

I hope this message finds you well. I wanted to share an update on the recent analysis we've conducted across three datasets (Users, Transactions, Products) and highlight some key findings, data quality issues and a notable trend we've identified.

Key Data Quality Issues:

1. Missing data: Several fields, particularly in the Products dataset, have significant number of missing values. For instance, the Category, Manufacturer, and Brand fields have large portions of missing entries.

2. Inconsistent Data Types: I have noticed that numerical fields like Total Sale and Total Quantity have string datatype and the data in Barcode column is considered as float instead of 'String'.

Interesting Trend:

One compelling insight I've observed relates to user age demographics. Among users aged 21 and over, the top five brands by number of receipts scanned are dominated by health and wellness related products, indicating a strong preference for these categories among this age group.

Outstanding Questions:

- 1) How should we handle the missing data in fields like Manufacturer and Category? Should we look into imputation (or) are these records incomplete and not essential for analysis?
- 2) What can be the expected format for fields like Barcode, which currently have inconsistent datatypes?

Request of Action:

Data validation: We could benefit from a review of the product and transaction data sources to classify whether missing fields are expected or if there's an issue.

Business Context: Any insights from your end on the significance of Manufacturer and Brand data for the analysis would be helpful, particularly if there are critical to product Categorization.

Let me know if you need more details or would like to discuss next steps. Your guidance on how to address the missing or inconsistent data would be invaluable as we continue our analysis.

Best regards,

Manasa Kolhepete

Data Analyst