Communication with Stakeholders

Hi Team,

As part of our work to build a reliable and scalable analytics layer on top of the receipts, users, brands, and item-level data, I ran an initial data quality assessment with the help of Python to ensure we are working with clean, consistent inputs. I wanted to share a few key findings and highlight where we might need alignment or support moving forward.

Across four main datasets (users, receipts, and brands), I identified a number of **data quality concerns** that could impact business insights, analytics accuracy, and downstream modeling:

Receipts:

- A sizable portion (about 40%) of receipts are missing core fields like purchase date, item list, or total spent — all of which are important for basic transaction-level analysis.
- There are also over 500 receipts where the pointsEarned field contains unexpected values like text ("N/A") or blanks instead of numbers.
- Over half of the items are missing barcodes. Without these, we can't link items to brands, which really limits our ability to do anything product- or brand-specific.

Brands:

- The brand dataset is generally in better shape, though about 13% of entries are missing category information, which makes classification tough.

Users:

- A subset of users have invalid or missing state data, and there are quite a few duplicate user IDs in the system — might just be test users or ingestion artifacts, but worth confirming.
- There are missing and duplicate data in the _id column. There is also discrepancy in the count of userId in the receipts data and the count of userId in the users data.

A lightweight data dictionary or list of field-level expectations would go a long way here — especially clarifying which fields are required, optional, or nullable under certain conditions. **There was some ambiguity that I'd love to clarify so we can clean this up**:

- Are barcodes always expected to be present, or are there known cases (e.g., manual entries) where they're intentionally left out?
- How should we treat user-flagged data that differs from scanned values (e.g., userFlaggedPrice)?
- Are receipts without item lists still considered valid transactions?

Performance and Scaling:

- We'll need to optimize joins between receipts, users, and items to support faster dashboarding and modeling — especially if we move to real-time insights.
- It might make sense to enforce some basic schema checks at the point of ingestion (e.g., barcode format, valid states, positive pricing).
- As new fields get introduced, a change log or schema registry could help keep analytics and product teams in sync.

Let me know if it'd be helpful to walk through this in more detail or flag anything for prioritization. I'm happy to adapt based on where this fits into upcoming product or reporting goals.

Thanks,

Lokesh