Subject: Data Quality Insights and Questions

Hi team,

I hope you are well.

I've been working on cleaning and analyzing the Fetch Rewards JSON datasets (‘users’, ‘brands’, and ‘receipts’) and wanted to share my findings and get some feedback. For more efficient data processing and analysis, I converted the unstructured data into a structured relational data model with four final tables using Python. The attached ER diagram details the relationships between the tables.

Key Issues:

- **Missing values:** High missing rates in the **Receipts** (39%-40% missing in critical fields) and **itemList** (more than half fields with over 80% missing) tables could affect analysis accuracy and completeness.
- **Item-brand linkage issue:** Over 55% of items in **itemList** table lack barcode data. Among available barcodes, only 3% (16 out of 552 unique barcodes) can be matched with the **brands** table, severely limiting brand level analysis.
- **User-receipts linkage issue**: Out of 1,119 unique receipts from 258 users, 148 receipts (13%) and 117 users (45%) cannot be mapped in the **Users** table.
- **Outliers:** There are abnormal totalSpent and quantityPurchased values (over $4000 and 600+ items per receipt).
- **Potential reward exploitation:** Some receipts show high bonus points for purchases under $10 with one or two items, indicating possible system gaming.

Questions:

- Missing values in **Receipts** table mainly come from receipts under status of ‘SUBMITTED’. Are 'SUBMITTED' receipts pending processing?
- Are there known barcode collection issues, and are there alternative fields for brand mapping?
- Is there always a delay in syncing between **receipts** and **users** tables that results into low user matching rate between them? If so, it should be taken into consideration for any customer-level analysis.
- Should we address extreme totalSpent and quantityPurchased values as outliers?
- Do we have fraud prevention mechanisms for unusual reward allocations (e.g., high bonus points with low spend)? If not, should we implement rules to mitigate system abuse?

- As the dataset grows, are we considering big data solutions (e.g., Hadoop, Spark) for more efficient processing and even real-time analytics?

Let me know your thoughts. I know it’s a lot of information to digest. I am happy to jump on a call to discuss with the team if needed. Thanks!

Best,

Sicheng Shen