**Communicate with Stakeholders**

Hi All,

I hope you're doing well. As I work with the Fetch Rewards Coding Exercise dataset, I’ve identified some potential data quality issues that could impact our insights and decision-making. To ensure we’re building reliable data assets, I’d love to get your input on a few key areas:

1. Data Clarity: Are there any specific guidelines or business rules around how key fields (e.g., totalSpent, pointsEarned, rewardsReceiptStatus) should be structured and validated?

2. Data Quality Issues: Through exploratory analysis, I’ve noticed missing values, duplicated records, inconsistent formats, and potential outliers in some key columns. For example, in Users data schema, the \_id column should have distinct values, but we see duplicated id numbers in the schema. Could you provide context on how this data is collected and whether these issues are expected?

3. Resolving Issues: To clean and optimize the data, I’d like to understand how these fields are used in decision-making. For example, in the question the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month, we see a variety of brands in the previous month, but we only see a few brands in the most recent month. Are there specific thresholds or business rules that should guide our corrections?

4. Additional Data Needs: Are there other data sources or attributes we should incorporate to enhance the accuracy and completeness of this dataset?

5. Performance & Scaling: As we prepare for production, I want to ensure our approach scales effectively. Do you foresee any constraints or challenges related to data volume, storage, or query performance that we should proactively address?

Your insights will help refine our data strategy and ensure we're aligned with business needs. Let me know if you’d like to discuss this further—I’m happy to set up a quick chat.

Looking forward to your thoughts!

Best,

Yangzhe