As part of our data analysis efforts, we’ve identified several data quality issues that require attention. I wanted to share these findings, discuss their implications, and outline steps to address them.

Key Observations:

1. Missing and Incomplete Data
   * Several fields contain null or incomplete values. While not all null values signify errors, nulls in unique identifier fields pose significant challenges
   * The `barcode` field appears unreliable as a key for matching items to receipts.
   * The `items` table lacks transactions from December 2020, though data is present for October and November 2020, as well as January and February 2021.
2. Duplicated Data
   * Some tables contain duplicate entries, such as multiple occurrences of the same `user\_id`
3. Prescence of Dummy Data:
   * Test or dummy data is present in various fields, affecting data reliability.
4. Inconsistent Data
   * Discrepancies exist between the total spend and total quantity purchased fields in `receipts ` and the aggregated values in the `items` table.

Questions to Address:

* Are there any known processes or business logic that might explain these inconsistencies (e.g., adjustments made outside the system or further downstream)?
* Can we confirm if test or dummy data is flagged in some way, or if it should be removed entirely?
* Are there additional data sources we can leverage to fill gaps (e.g., external transaction logs or audit trails)?
* What is the current method or algorithm used to define `brandCode`?
* Is there a data dictionary for the items table (`receipts.rewardsReceiptItemList`)?

Next Steps:

* Define the expected behavior and validation rules for key metrics, such as total spend and item quantities, to ensure accurate aggregations.
* Confirm whether barcode is the intended key for linking items to receipts. If not, identify a reliable alternative. We identified brandCode as a potential alternative and applied partial string matching techniques to enhance the data availability and completeness.
* Verify the completeness of the December 2020 data—was it not recorded, or is it stored elsewhere?
* Access to any existing documentation about data pipelines or ETL processes to better understand data flow and transformations.

Performance and Scaling Considerations:

As we optimize these datasets for analysis, we anticipate the following challenges in production:

1. Data Volume: Large tables like items and receipts may grow exponentially, impacting query performance.
2. Joins and Aggregations: Complex joins and aggregations may slow down reporting as the database scales.

Proposed Solutions:

1. Implement indexing for commonly queried fields like `receipt\_id` and `user\_id`.
2. Introduce data partitioning by date to manage growing volumes.
3. Monitor query performance and optimize SQL scripts as needed.

Let me know if you’d like to discuss this further or if additional context would be helpful.