Hello,

I hope this message finds you well. I have been analyzing our users, receipts, and brands data, and I've identified some key questions and data quality issues that need addressing:

**Questions About the Data:**

Users Data Schema: Why do we have multiple records with the same user IDs and timestamps?

Brands Data Schema: Could you provide more context about the CPG collection? Is there another table schema for CPG-related data?

**Data Quality Issues Discovered:**

I used various Python libraries and functions to uncover these issues:

1. isnull() to identify null values
2. duplicated() to check for duplicate rows
3. unique() to list unique values in a column
4. value\_counts() to calculate the count of categories
5. matplotlib.pyplot for visualizing categorical variable distributions
6. isin() to validate values in variables
7. Checks for negative or unrealistic values
8. Ensured referential integrity between tables
9. describe() to generate summary statistics and identify outliers
10. Checked the logical consistency of date variables

**Some pressing data quality issues include:**

Receipts Data:

1. Outliers in the purchasedItemCount variable (min = 0, max = 689, mean ≈ 14)
2. Numerous missing values in bonusPointsEarned, bonusPointsEarnedReason, finishedDate, pointsAwardedDate, pointsEarned, purchaseDate, purchasedItemCount, and totalSpent.

Users Data:

1. Over half of the records are duplicates (283 out of 495)
2. Missing values in lastLogin, signUpSource, and state
3. signUpSource variable shows 96.23% of users signed up via email, 1.42% via Google
4. 91.04% of users are from Wisconsin.

Brands Data:

1. Variables like ‘brandCode’, ‘category’, ‘categoryCode’ and ‘topBrand’ have many missing values
2. % of missing values for brandCode : 20.05%, category : 13.28%, categoryCode : 5.69% and topBrand : 52.44%
3. 31 of 555 brands are considered as top brands
4. ‘Baking’ category takes up 31.62% of all categories

**Information Needed to Resolve Data Quality Issues:**

1. The rewardsReceiptItemList variable in the receipts/items data schema has multiple null fields and inconsistencies.
2. Clarification on vague variables like originalMetaBriteBarcode.
3. Barcodes in rewardsReceiptItemList not present in the brands table.
4. Some items have duplicate barcodes.

* **What other information would you need to help you optimize the data assets you're trying to create?**

1. It would be beneficial to know the primary metrics and KPIs of the business like total spend, user engagement, etc.
2. The frequency of data updates in brands, users and receipts tables
3. The way missing or incorrect values are handled
4. The date ranges for which data is available in tables
5. Current policies regarding who can access and modify the data
6. Which data is considered PII and way sensitive data is handled

* **What performance and scaling concerns do you anticipate in production and how do you plan to address them?**

1. Growing receipts data table can cause issues in future due to ‘rewardsReceiptItemList’ variable which contains multiple items in the list. This can get complicated and slow down the performance. It can be rectified by creating another table for items and store information regarding items separately.
2. Concurrent access by multiple users can lead to locking issues and reduced performance. This can be resolved by ensuring proper opening, committing and closing of cursors while accessing database tables and implementing proper load balancing.
3. Ensuring referential integrity and dealing with high insert/update rates can be challenging. Use database constraints (e.g., foreign keys) and implement data validation checks
4. Scaling the database infrastructure to handle increasing loads and data volume. Use a scalable database solution e.g., Amazon RDS, Google Cloud SQL.