**Subject: Data Analysis Findings and Recommendations**

Hi,

As part of the recent analysis, I reviewed the available datasets and uncovered some data quality issues, identified interesting trends, and have outlined actionable recommendations. Below is a summary of my findings:

**Key Data Quality Issues and Outstanding Questions**

1. **Data Quality Issues**:
   * **users\_df**: Missing values in key fields:
     + BIRTH\_DATE: 3,675 missing values.
     + STATE: 4,812 missing values.
     + LANGUAGE: 30,508 missing values.
     + GENDER: 5,893 missing values.
   * **transactions\_df**: 5,762 missing values in the BARCODE column.
   * **products\_df**: Missing values in several important columns:
     + CATEGORY\_1: 111 missing values.
     + CATEGORY\_2: 1,424 missing values.
     + CATEGORY\_3: 60,566 missing values.
     + CATEGORY\_4: 778,093 missing values.
     + MANUFACTURER: 226,474 missing values.
     + BRAND: 226,472 missing values.
2. **Outstanding Questions**:
   * What do the CATEGORY\_1, CATEGORY\_2, etc., columns in products\_df represent? These column names are unclear and may need renaming for better interpretability.
   * Should missing values in critical fields (e.g., BIRTH\_DATE, BRAND, BARCODE) be replaced, inferred, or removed?

**Interesting Trend**

One key finding from the analysis is the dominance of specific brands in user engagement and sales:

* **Top 5 Brands by Receipts Scanned (Users 21 and Over)**:
  1. COCA-COLA: 628 receipts scanned.
  2. ANNIE'S HOMEGROWN GROCERY: 576 receipts scanned.
  3. DOVE: 558 receipts scanned.
  4. BAREFOOT: 552 receipts scanned.
  5. ORIBE: 504 receipts scanned.
* **Top 5 Brands by Sales (Users with Accounts ≥6 Months)**:
  1. COCA-COLA: $2,592.10.
  2. ANNIE'S HOMEGROWN GROCERY: $2,383.92.
  3. DOVE: $2,327.47.
  4. BAREFOOT: $2,284.59.
  5. ORIBE: $2,085.93.

This highlights that COCA-COLA and ANNIE'S HOMEGROWN GROCERY consistently outperform other brands in both receipts and sales, indicating strong user loyalty and brand presence.

**Request for Action**

To move forward, I recommend:

1. Addressing missing values in key fields, particularly BIRTH\_DATE, BARCODE, and BRAND.
2. Clarifying the definitions of the CATEGORY\_1, CATEGORY\_2, etc., columns in products\_df to ensure proper usage.
3. Defining "power users" more comprehensively by considering additional metrics like frequency or recency of transactions.

Please let me know if we can coordinate with relevant teams or obtain further input to address the data quality issues and improve our analysis.

Best regards,  
Yashaswini Ganne  
Data Analyst