Hi Team,

I wanted to share the results of my recent investigation into the data quality and trends we've observed in our datasets. Below is a summary:

Key Data Quality Issues:

1. **Transaction Dataset:**
    - Missing values in FINAL_SALE and BARCODE make it challenging to analyze total receipts or product details accurately.
    - Some records show inconsistencies where the SCAN_DATE occurs before the PURCHASE_DATE.
2. **User Dataset:**
    - Missing information in BIRTH_DATE, GENDER, and STATE. While this may not significantly affect overall analysis, having complete profiles would improve customer segmentation efforts.
3. **Products Dataset:**
    - 0.4% of barcodes are missing, and there are duplicate barcodes with conflicting category information. Barcodes are critical for accurately identifying products and linking them to transactions.

Interesting Trend:

We observed that users aged 21 and older predominantly purchase products from brands such as coca-cola, annie’s homegrown grocery, and Dove. These brands could potentially be key partners for targeted promotions or partnerships.

Request for Action:

1. **Clarifications Needed:**
    - What do "zero" values in FINAL_QUANTITY and FINAL_SALE represent? Are they placeholders or legitimate data?
    - Are the duplicate barcodes reflective of real product overlaps, or do we need to standardize this information?
2. **Support Required:**
    - Input from the product team on the hierarchy and purpose of the CATEGORY_1 through CATEGORY_4 fields.

- Assistance from the engineering team to investigate the source of date discrepancies and missing values in the datasets.

By addressing these issues, we can ensure better data quality and uncover more actionable insights for the business. Please let me know if we can meet to discuss these points further or if you can connect me with the relevant teams.

Best regards,

Ariel Lai