**First: explore the data**

**Review the unstructured csv files and answer the following questions with code that supports your conclusions:**

* Are there any data quality issues present?
* Are there any fields that are challenging to understand?

**We recommend using SQL or Python and data visualization to examine the data.**

Are there any data quality issues present?  
  
**Data Quality Issues:**

1. **Transactions Table:**
   * The BRAND column contains many null values and entries labelled **"BRAND NOT KNOWN"**, which should be cleaned or standardized.
   * The BARCODE column has missing values (nulls), affecting product identification.
2. **Users Table:**
   * Several key fields have missing values:
     + BIRTH\_DATE (3675 missing)
     + STATE (missing, not important to us, so it can be ignored)
     + LANGUAGE (missing, not important to us, so it can be ignored)
     + GENDER (missing, not important to us, so it can be ignored)
   * These gaps in Birthday impacts finding the outputs where age and generations are involved.
3. **Products Table:**
   * The CATEGORY\_4 column is sparsely populated (null values), making it unreliable for categorization.
   * The CATEGORY\_3 column has missing values, but the CATEGORY\_1 column remains relatively complete, making it a better choice for classification.
   * The MANUFACTURER and BRAND columns contain missing values, which could affect product insights.
   * The BARCODE field has missing values, which may create inconsistencies when linking products with transactions.

Are there any fields that are challenging to understand?

**Final Quantity (FINAL\_QUANTITY)**

* There are instances where the value is "zero", but a price is still recorded in FINAL\_SALE. This raises questions about the transaction's accuracy.

**Possible Explanations for This Issue:**

* **Coupon or Discount Adjustments:** It might indicate an item was refunded or discounted, but the system still registered a price adjustment.
* **Machine or Data Entry Error:** A technical issue could have caused the system to log the price despite the quantity being zero.
* **Promotional Free Items:** Some stores register promotional items as having zero quantity, but their price is still shown (e.g., "Buy One Get One Free" offers).
* **Refund or Reversal Issue:** If a return was processed incorrectly, the price might still be reflected while the quantity is set to zero.

**How to Fix It?(Need to check with stakeholder)**

* Investigate whether these transactions correspond to a refund, discount, or promo event.
* Check for store policies on free items or discounts to validate the logic.
* Standardize how refunds or promotions are recorded to ensure data consistency.

**Second: provide SQL queries**

**Answer three of the following questions with at least one question coming from the closed-ended and one from the open-ended question set. Each question should be answered using one query.**

**Closed-ended questions:**

1. What are the top 5 brands by receipts scanned among users 21 and over?

A screenshot of a computer

AI-generated content may be incorrect.

**Code:**

**--1. Top 5 brands by receipts scanned among users 21 and over**

*WITH UserAge AS (*

*SELECT ID,*

*(STRFTIME('%Y', 'now') - STRFTIME('%Y', BIRTH\_DATE)) AS AGE*

*FROM USER\_TAKEHOME*

*WHERE (STRFTIME('%Y', 'now') - STRFTIME('%Y', BIRTH\_DATE)) >= 21*

*)*

*SELECT pt.BRAND, COUNT(tt.RECEIPT\_ID) AS RECEIPT\_COUNT*

*FROM TRANSACTION\_TAKEHOME tt*

*JOIN UserAge ua ON tt.USER\_ID = ua.ID*

*JOIN PRODUCTS\_TAKEHOME pt ON tt.BARCODE = pt.BARCODE*

*WHERE pt.BRAND IS NOT NULL AND pt.BRAND <> ''*

*GROUP BY pt.BRAND*

*ORDER BY RECEIPT\_COUNT DESC*

*Limit 5;*

1. What are the top 5 brands by sales among users that have had their account for at least six months?

A screenshot of a menu

AI-generated content may be incorrect.

**Code:**

**--2. Top 5 brands by sales among users with accounts for at least 6 months**

*WITH ActiveUsers AS (*

*SELECT ID*

*FROM USER\_TAKEHOME*

*WHERE CREATED\_DATE <= DATE('now', '-6 months')*

*)*

*SELECT pt.BRAND, SUM(tt.FINAL\_SALE) AS TOTAL\_SALES*

*FROM TRANSACTION\_TAKEHOME tt*

*JOIN ActiveUsers au ON tt.USER\_ID = au.ID*

*JOIN PRODUCTS\_TAKEHOME pt ON tt.BARCODE = pt.BARCODE*

*WHERE pt.BRAND IS NOT NULL AND pt.BRAND <> ''*

*GROUP BY pt.BRAND*

*ORDER BY TOTAL\_SALES DESC*

*LIMIT 5;*

**Open-ended questions: for these, make assumptions and clearly state them when answering the question.**

1. Which is the leading brand in the Dips & Salsa category?

**Code:**

*SELECT pt.BRAND, SUM(tt.FINAL\_SALE) AS TOTAL\_SALES*

*FROM TRANSACTION\_TAKEHOME AS tt*

*JOIN PRODUCTS\_TAKEHOME AS pt ON tt.BARCODE = pt.BARCODE*

*WHERE pt.CATEGORY\_2 LIKE 'Dips & Salsa' and pt.BRAND IS NOT NULL AND pt.BRAND <> ''*

*GROUP BY pt.BRAND*

*ORDER BY total\_sales DESC*

**Explanation:**

Based on the assumptions, we have considered Sales to figure out leading brand in Dips and Salsa and we figured out that TOSTITOS is leading with a total sales of 103,354.84.



**Third: communicate with stakeholders**

**Construct an email or slack message that is understandable to a product or business leader who is not familiar with your day-to-day work. Summarize the results of your investigation. Include:**

**Mail to stakeholders**

Hi Everyone,

I hope you're doing well. I wanted to share a summary of the results from my recent investigation into the data, along with some key findings and questions that need further clarification.

**Key Data Quality Issues & Outstanding Questions:**

* **BARCODE Matching Issue**: There appears to be a mismatch between the BARCODE values in the TRANSACTION\_TAKEHOME and PRODUCTS\_TAKEHOME tables, which could cause join issues during analysis.
* **Timeframe Definition**: We need clarity on the analysis timeframe—should it include all available data, or should we focus only on the last 12 months?
* **Zero Final Quantity Values**: I've encountered two values for a single transaction where the Final Quantity is either zero or one. It seems that the zero value is storing the total value of the transaction, which could lead to discrepancies in the analysis.

**Interesting Trend in the Data:**

One interesting trend I found is that **long-term users (>12 months)** tend to spend more on **premium brands** (e.g., Dove, Barefoot), while **new users (<6 months)** are more likely to purchase **budget-friendly brands**. This suggests that brands targeting long-term customers could benefit from loyalty programs, while budget brands may consider offering discounts to attract newer users.

**Request for Action:**

To help resolve some of the data quality issues and fine-tune the analysis, I need some additional information:

* Are there specific **date ranges** we should focus on for trend analysis?
* Should we include **pending transactions**, or only completed ones in the analysis?
* Does "sales" refer to the FINAL\_SALE (total revenue per transaction), or another metric?
* There were errors related to missing columns (CATEGORY\_2, DOB\_Year). Could you confirm if we can compute those missing values, or if they are essential for our analysis?

Your help in answering these questions will be crucial for finalizing the analysis and ensuring that we draw actionable insights.

Looking forward to your feedback!

Best regards,  
Aishwarya Kalleshmurthy.