Subject: Data Quality Issues and Insights from Recent Analysis

Hi [Stakeholder Name],

I wanted to provide a summary of the recent data investigation, including key findings and a few questions that need clarification.

Key Data Quality Issues:

1. **Missing Data**:
    - In the **Product Table**, the fourth category level (**CATEGORY_4**) is missing for most records. Since the first three category levels provide enough detail, I recommend dropping **CATEGORY_4** to streamline the data.
    - In the **Transaction Table**, 5,000 entries are missing **BARCODE** information, which makes it impossible to trace these transactions back to specific products. I plan to drop these rows for accurate analysis,
2. **Duplicate Receipts**:
    - There are instances where **RECEIPT_ID** (meant to uniquely identify each transaction) appears multiple times. To ensure accuracy, I plan to retain only the unique transactions where both **FINAL_SALE** and **FINAL_QUANTITY** are non-zero.
3. **Unusual Records**:
    - A notable issue is 12,500 records where **FINAL_QUANTITY** is zero, but **FINAL_SALE** still has a value. This may reflect anomalies like returns, bulk discounts, or data entry errors. I'm holding off on next steps until we better understand this situation.
4. **User Data**:
    - In the **User Table**, several fields, such as **BIRTH_DATE**, **STATE**, and **GENDER**, are missing or contain placeholder values. To retain all users linked to transactions, I suggest filling these missing fields with "unknown" rather than dropping the records.

Key Insights:

- **Category Sales Trends**: Florida customers have the highest overall sales. Among product categories, **Snacks** leads, followed by **Health & Wellness** and **Beverages**. In Snacks, **Candy** and **Chips** are the top sellers, while **Carbonated**

Soft Drinks make up 94% of Beverage sales over the last four months.

- **Brand Leadership**: In the **Dips & Salsa** category, **TOSTITOS** is the clear leader, consistently outperforming other brands in both sales and quantity sold. Sales were particularly high in June and July, suggesting a summer spike. Additionally, 46% of TOSTITOS sales came from Walmart, hinting at potential for targeted promotions during this period.
- **Store Leadership**: **Walmart** accounted for 41% of total sales in the past four months, consistently outperforming other stores each month. Walmart stores in Florida, South Carolina, and Texas had particularly high sales. These regions might benefit from increased promotional efforts to drive further engagement.

Request for Action:

1. **Clarification on Duplicate Receipts**: Could you clarify whether duplicate **RECEIPT_IDs** are intentional (e.g., multiple items in one transaction) or if this is a data entry issue? This will help us decide how to handle the 12,500 rows where **FINAL_QUANTITY** is zero but **FINAL_SALE** is not.
2. **Further Explanation of Final Quantity vs. Final Sale**: I need more information about why there are instances where **FINAL_QUANTITY** is zero, but **FINAL_SALE** has a value. Is this due to a special case in our transaction process (e.g., partial refunds, promotional items), or should it be treated as an error?
3. **Additional User Information**: It would be helpful to have more demographic information about users, such as age. This could enable us to create better customer profiles and understand if top-selling brands like TOSTITOS have higher sales in specific age groups or states.

Thanks in advance for your insights. I'll follow up with further analysis once we resolve these outstanding questions.

Best regards,
Doris Zhang