Answer 1) **Initial Data Exploration Findings:**

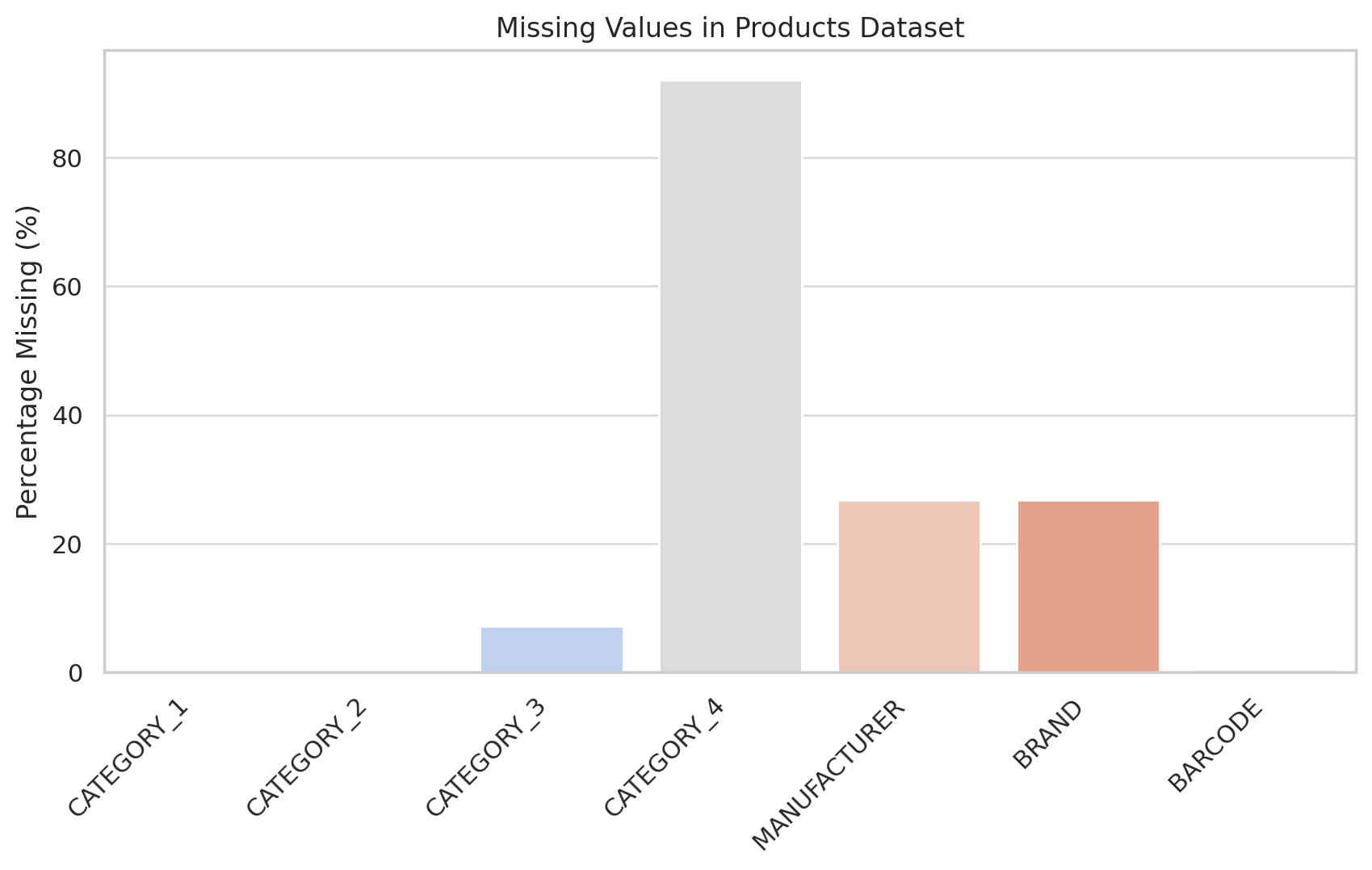
**1. Data Quality Issues:**

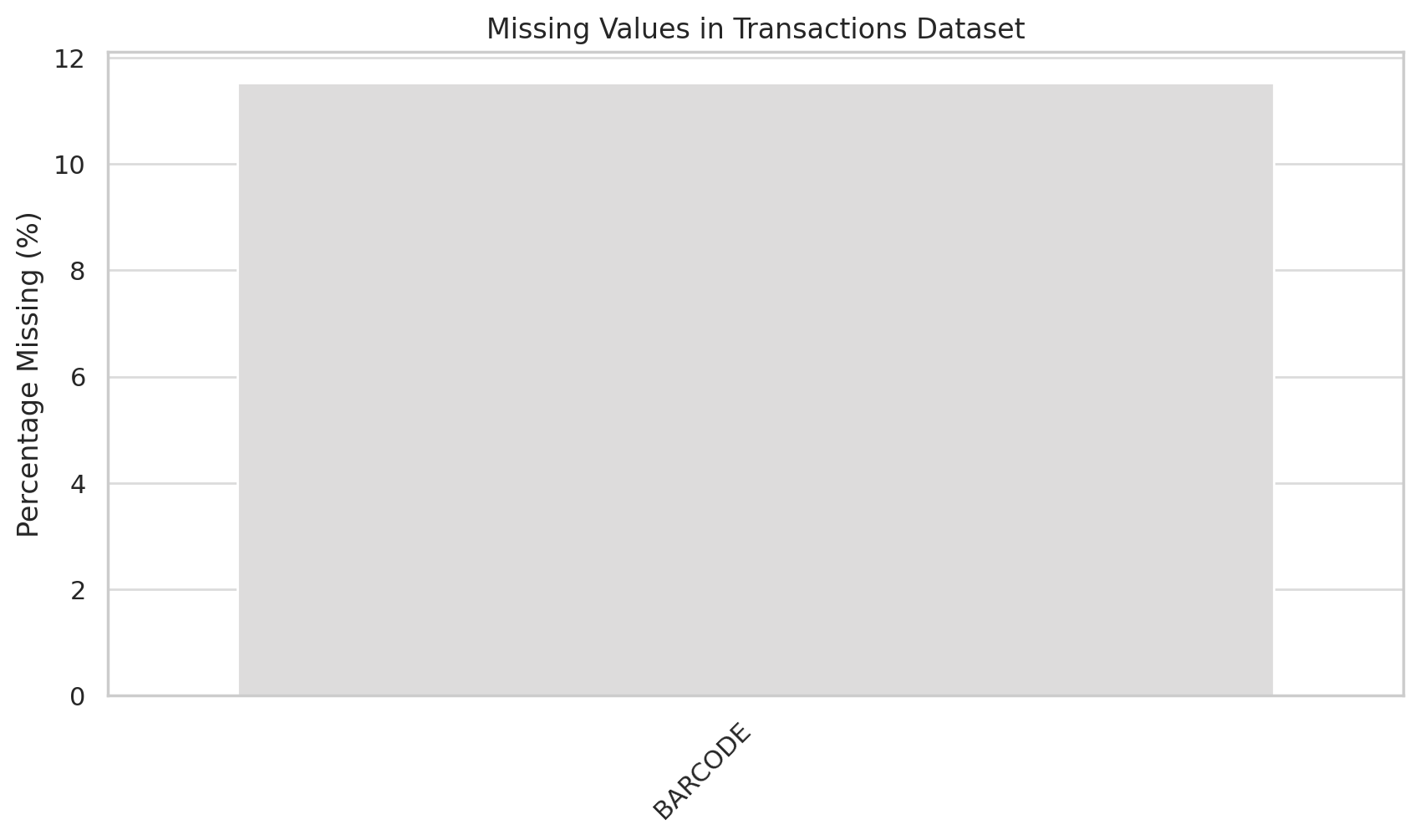
* **Missing Values:**
  + PRODUCTS\_TAKEHOME.csv:
    - CATEGORY\_4 has significant missing values (only 67,459 out of 845,552).
    - CATEGORY\_3, MANUFACTURER, and BRAND also have many missing values.
    - BARCODE has some missing values.
  + TRANSACTION\_TAKEHOME.csv:
    - BARCODE has missing values (~5,762 out of 50,000).
  + USER\_TAKEHOME.csv:
    - BIRTH\_DATE, STATE, LANGUAGE, and GENDER contain missing values.
* **Data Type Issues:**
  + BARCODE is stored as a float64, which may cause precision issues.
  + FINAL\_QUANTITY has string values like "zero" instead of numeric 0.
* **Potential Inconsistencies:**
  + FINAL\_SALE column in TRANSACTION\_TAKEHOME.csv sometimes appears empty.

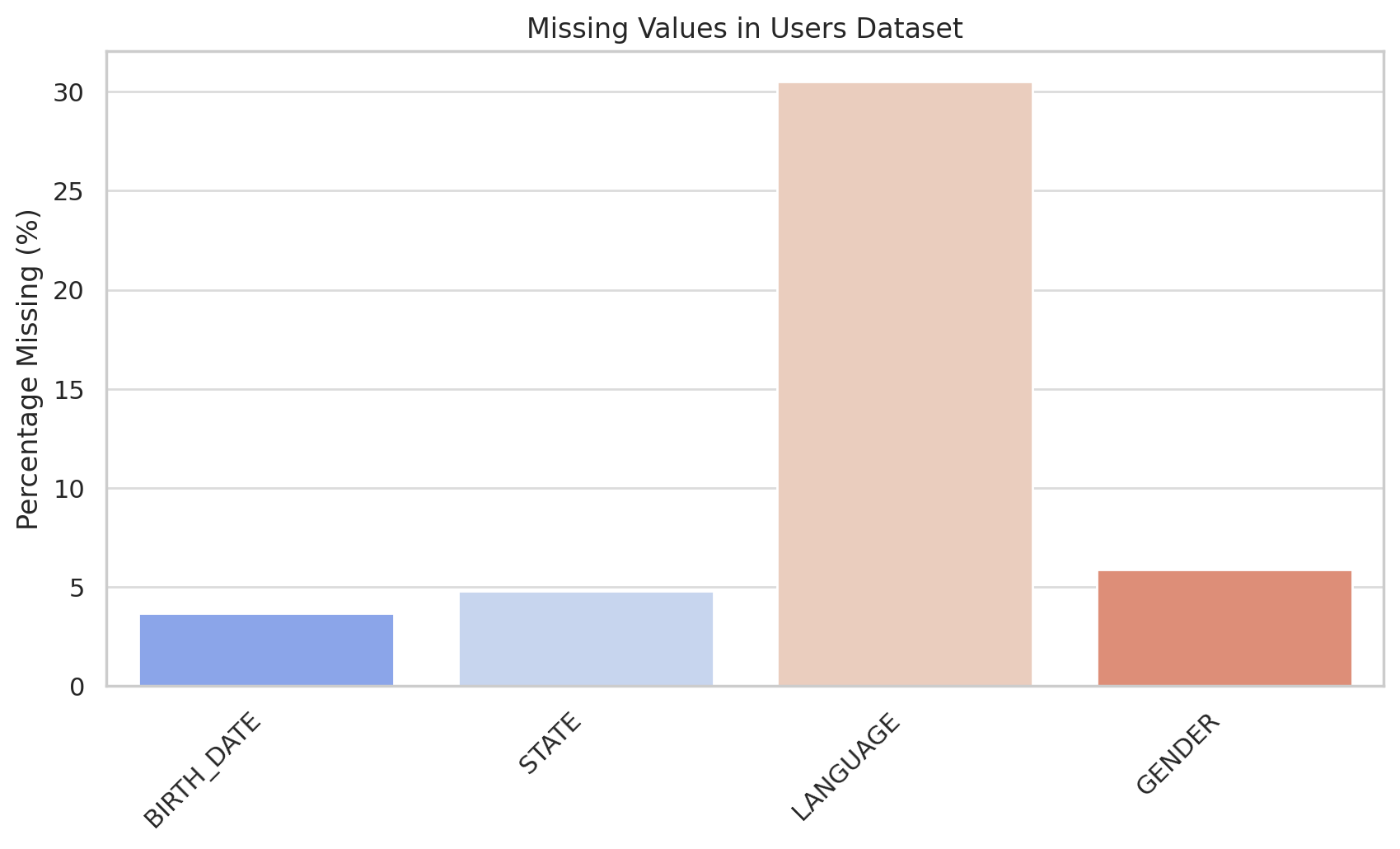
**2. Challenging Fields:**

* CATEGORY\_1 to CATEGORY\_4: Hierarchical structure may need clarification.
* LANGUAGE uses codes like es-419 (Latin American Spanish), which might not be immediately understandable.
* FINAL\_QUANTITY contains string representations of numbers ("zero") instead of numeric values.

Next, there is data summaries and visualizations to further explore these issues. ​​







The missing values visualization confirms that CATEGORY\_4, MANUFACTURER, and BRAND in the Products dataset have substantial gaps. The Transactions dataset has missing values in BARCODE, and the Users dataset has missing data in BIRTH\_DATE, STATE, LANGUAGE, and GENDER.

Next, I will check for inconsistencies in categorical fields and unusual data patterns. ​​

**Additional Findings:**

**1. Products Dataset:**

* **Hierarchical Complexity:**
  + CATEGORY\_1 has 27 unique values, while CATEGORY\_4 has 127, indicating a deep hierarchical categorization.
  + The number of manufacturers (4,354) and brands (8,122) suggests a highly diverse dataset, potentially leading to inconsistencies in naming conventions.

**2. Transactions Dataset:**

* **Store Distribution:**
  + WALMART dominates the transactions (21,326 entries), which could impact analysis if not normalized.
* **Quantity Issues:**
  + FINAL\_QUANTITY contains "zero" instead of 0, indicating a need for data cleaning.
  + There are 87 unique values, suggesting possible non-integer values (e.g., 6.22, 1.22), which might be unexpected.

**3. Users Dataset:**

* **Language Codes:**
  + Most users speak en (English) or es-419 (Latin American Spanish).
* **Gender Categories:**
  + There are multiple variations for similar categories (Non-Binary vs. non\_binary, "My gender isn't listed" as a unique entry), suggesting standardization is needed.

Answer 2) Closed-ended question: Top 5 brands by receipts scanned among users 21 and over

SELECT p.BRAND, COUNT(t.TRANSACTION\_ID) AS receipt\_count

FROM TRANSACTION\_TAKEHOME t

JOIN PRODUCTS\_TAKEHOME p ON t.BARCODE = p.BARCODE

JOIN USER\_TAKEHOME u ON t.USER\_ID = u.USER\_ID

WHERE TIMESTAMPDIFF(YEAR, u.BIRTH\_DATE, CURDATE()) >= 21

GROUP BY p.BRAND

ORDER BY receipt\_count DESC

LIMIT 5;

-- Open-ended question: Identifying Fetch's power users

-- Assumption: Power users are those who have scanned receipts most frequently in the past year.

SELECT u.USER\_ID, COUNT(t.TRANSACTION\_ID) AS total\_receipts, SUM(t.FINAL\_SALE) AS total\_spent

FROM TRANSACTION\_TAKEHOME t

JOIN USER\_TAKEHOME u ON t.USER\_ID = u.USER\_ID

WHERE t.TRANSACTION\_DATE >= DATE\_SUB(CURDATE(), INTERVAL 1 YEAR)

GROUP BY u.USER\_ID

ORDER BY total\_receipts DESC

LIMIT 10;

-- Open-ended question: Leading brand in the Dips & Salsa category

-- Assumption: "CATEGORY\_2" contains "Dips & Salsa" for relevant products.

SELECT p.BRAND, SUM(t.FINAL\_SALE) AS total\_sales

FROM TRANSACTION\_TAKEHOME t

JOIN PRODUCTS\_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE p.CATEGORY\_2 = 'Dips & Salsa'

GROUP BY p.BRAND

ORDER BY total\_sales DESC

LIMIT 1;

Answer 3)

**Subject:** Data Quality Issues & Key Insights from Fetch Analysis

Hi Sir/Ma’am,

We've conducted an initial review of Fetch’s transaction, product, and user data. Below are key findings and next steps:

**Data Quality Issues:**

* Missing data in key fields such as CATEGORY\_4, MANUFACTURER, BARCODE, and BIRTH\_DATE, which could impact analysis accuracy.
* Inconsistent values in FINAL\_QUANTITY (e.g., "zero" instead of numeric 0) and GENDER (e.g., multiple variations for non-binary users).
* Potential misclassification in hierarchical categories (CATEGORY\_1 to CATEGORY\_4).

**Key Trend:**

* Walmart dominates transactions, accounting for over 40% of scanned receipts. This concentration might skew insights unless weighted appropriately.

**Next Steps & Request for Input:**

1. Can we confirm if missing CATEGORY\_4 values should be inferred from related fields?
2. Should FINAL\_QUANTITY and GENDER be standardized to improve consistency?
3. Do we need to adjust analysis methods to account for Walmart's dominance?

Looking forward to your thoughts.

Best Regards,  
Pushpanjali Chauhan