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

SELECT brand\_name, COUNT(receipt\_id) AS receipt\_count

FROM receipts

JOIN users ON receipts.user\_id = users.user\_id

WHERE users.age >= 21

GROUP BY brand\_name

ORDER BY receipt\_count DESC

LIMIT 5;

-- Open-ended question: Who are Fetch’s power users?

-- Assumption: Power users are those who scan the most receipts within the past 6 months

SELECT user\_id, COUNT(receipt\_id) AS total\_receipts

FROM receipts

WHERE scan\_date >= DATE\_SUB(CURRENT\_DATE, INTERVAL 6 MONTH)

GROUP BY user\_id

ORDER BY total\_receipts DESC

LIMIT 10;

-- Additional query: Percentage of sales in the Health & Wellness category by generation

SELECT users.generation,

SUM(receipts.sales\_amount) \* 100.0 / (SELECT SUM(sales\_amount) FROM receipts WHERE category = 'Health & Wellness') AS sales\_percentage

FROM receipts

JOIN users ON receipts.user\_id = users.user\_id

WHERE receipts.category = 'Health & Wellness'

GROUP BY users.generation;

Answer 1) # Visualizing missing values

plt.figure(figsize=(10, 6))

sns.heatmap(df.isnull(), cmap='viridis', cbar=False)

plt.title("Missing Values Heatmap")

plt.show()

# Identifying challenging fields based on inconsistent formats

def check\_inconsistent\_values(df, name):

print(f"\nChecking for inconsistencies in: {name}\n")

for col in df.columns:

unique\_vals = df[col].dropna().unique()

if len(unique\_vals) < 20:

print(f"Column: {col}, Unique Values: {unique\_vals}")

elif df[col].dtype == 'object':

print(f"Column: {col}, Sample Values: {df[col].dropna().sample(5).values}")

for name, df in dataframes.items():

check\_inconsistent\_values(df, name)

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, I'll generate 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)