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. I

Part 1: Data Exploration

Are there any data quality issues present?

Yes — a few notable ones showed up while exploring the datasets:

- **PRODUCTS_TAKEHOME.csv**
    - The CATEGORY_4 field is missing for 92% of the rows — seems like it might be optional or not widely used.
    - Around 27% of the rows are missing both MANUFACTURER and BRAND, which could limit any brand-level analysis.
    - CATEGORY_3 has an unusually high number of unique values (~60K+), which makes me think there might be some messy or overly granular labeling.
    - A small number of rows (about 0.5%) are missing BARCODE, which could cause issues when joining with transaction data.
- **USER_TAKEHOME.csv**
    - Fields like BIRTH_DATE, STATE, and especially LANGUAGE have missing values (with LANGUAGE missing in 30% of users).
    - The GENDER field has inconsistent labeling — there are many variants like "non_binary", "Non-Binary", "Prefer not to say", etc. I'd definitely recommend normalizing these.
- **TRANSACTION_TAKEHOME.csv**
    - Some entries in FINAL_QUANTITY are listed as "zero" (string) instead of numeric 0, which might throw off calculations until cleaned.
    - FINAL_SALE also has blank string entries instead of nulls.
    - About 11.5% of rows are missing a BARCODE, which again impacts our ability to tie transactions back to specific products.

Any fields that are challenging to understand?

A few stood out:

- CATEGORY_3 has so many unique values that it's hard to know how to group or interpret them without some cleaning or mapping logic.
- FINAL_QUANTITY and FINAL_SALE both have inconsistent formats — mixing strings, blanks, and numbers.
- GENDER is especially messy and could easily be simplified into fewer categories for analysis.
- BARCODE appears in all datasets, but is often missing — making joins less reliable unless filtered.

Part 2: SQL-Based Questions

Closed-Ended Question:

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

Assumptions:

- I'm defining “account age” as the time between a user's CREATED_DATE and the PURCHASE_DATE of a receipt.
- Six months is approximated as **180 days**.
- Some sales values (FINAL_SALE) were blank or non-numeric, so I filtered those out before calculating brand sales.
- I only included transactions where BARCODE mapped to a known BRAND.

SQL Query:

```
WITH user_with_age AS (
  SELECT
    u.ID AS USER_ID,
    u.CREATED_DATE,
    t.PURCHASE_DATE,
    t.BARCODE,
    t.FINAL_SALE,
    DATE_PART('day', t.PURCHASE_DATE - u.CREATED_DATE) AS account_age_days
  FROM TRANSACTION_TAKEHOME t
  JOIN USER_TAKEHOME u ON t.USER_ID = u.ID
  WHERE t.PURCHASE_DATE IS NOT NULL AND u.CREATED_DATE IS NOT NULL
),
eligible_users AS (
   SELECT * FROM user_with_age
   WHERE account_age_days >= 180
),
joined_products AS (
   SELECT
      e.BARCODE,
      e.FINAL_SALE,
      p.BRAND
   FROM eligible_users e
   JOIN PRODUCTS_TAKEHOME p ON e.BARCODE = p.BARCODE
   WHERE e.FINAL_SALE IS NOT NULL AND p.BRAND IS NOT NULL
)

SELECT
   BRAND,
   ROUND(SUM(CAST(FINAL_SALE AS FLOAT)), 2) AS total_sales
FROM joined_products
GROUP BY BRAND
ORDER BY total_sales DESC
LIMIT 5;
```

Open-Ended Question:

What is the percentage of sales in the Health & Wellness category by generation?

Assumptions:

- Generations are based on **birth year** from BIRTH_DATE:
  - Gen Z: 1997+
  - Millennials: 1981–1996
  - Gen X: 1965–1980
  - Boomers: before 1965
- I only used users who had a valid BIRTH_DATE, and products with non-null CATEGORY_1.
- Sales in the Health & Wellness category are identified using CATEGORY_1 = 'Health & Wellness' in the products table.
- I excluded transactions where FINAL_SALE was null or blank.

SQL Query:

```
WITH user_generation AS (
   SELECT
      ID AS USER_ID,
      CASE
```

```
            WHEN DATE_PART('year', TO_DATE(BIRTH_DATE, 'YYYY-MM-DD')) >= 1997 THEN
'Gen Z'
            WHEN DATE_PART('year', TO_DATE(BIRTH_DATE, 'YYYY-MM-DD')) BETWEEN 1981
AND 1996 THEN 'Millennials'
            WHEN DATE_PART('year', TO_DATE(BIRTH_DATE, 'YYYY-MM-DD')) BETWEEN 1965
AND 1980 THEN 'Gen X'
            ELSE 'Boomers'
        END AS generation
    FROM USER_TAKEHOME
    WHERE BIRTH_DATE IS NOT NULL
),
filtered_tx AS (
    SELECT
        t.FINAL_SALE,
        t.USER_ID,
        p.CATEGORY_1
    FROM TRANSACTION_TAKEHOME t
    JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE
    WHERE t.FINAL_SALE IS NOT NULL AND p.CATEGORY_1 = 'Health & Wellness'
),
sales_by_gen AS (
    SELECT
        ug.generation,
        SUM(CAST(ft.FINAL_SALE AS FLOAT)) AS gen_sales
    FROM filtered_tx ft
    JOIN user_generation ug ON ft.USER_ID = ug.USER_ID
    GROUP BY ug.generation
),
total_sales AS (
    SELECT SUM(gen_sales) AS total FROM sales_by_gen
)

SELECT
    s.generation,
    ROUND((s.gen_sales / t.total) * 100, 2) AS health_wellness_sales_pct
FROM sales_by_gen s, total_sales t
ORDER BY health_wellness_sales_pct DESC;
```

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:

- Key data quality issues and outstanding questions about the data
- One interesting trend in the data
  - Use a finding from part 2 or come up with a new insight
- Request for action: explain what additional help, info, etc. you need to make sense of the data and resolve any outstanding issues

Subject: Summary of Initial Data Analysis

Hi [Name],

I've completed an initial review of the transaction, product, and user datasets. Here are a few key points:

- Data quality: Brand and manufacturer fields are missing in ~25% of product records; GENDER has inconsistent values; FINAL_QUANTITY and FINAL_SALE need standardization.
- Interesting trend: Millennials contribute to 46% of Health & Wellness category sales, indicating strong engagement with wellness products.
- Next steps: A data dictionary (especially for product categories) and clarification on whether each row represents a line item or full receipt would help deepen the analysis.

Happy to explore further based on your priorities.

Best,
Likhita