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?

  Yes there are Data Quality Issues

  USER_TAKEHOME.csv:

- **Data Types**: All fields are stored as object types, including date fields, which would benefit from conversion to datetime formats for easier analysis.
- **Missing Values**:
  - BIRTH_DATE has 3,675 missing values, which could impact age-related metrics.
  - STATE, LANGUAGE, and GENDER contain varying levels of missing data (4,812, 30,508, and 5,892 respectively).
- **Anomalies**:
  - GENDER contains 11 unique values, suggesting inconsistencies or possible typos in categorization. While "female" is the most frequent, some entries appear inconsistent.

  TRANSACTION_TAKEHOME.csv:

- **Data Types**: BARCODE is currently a float64; converting it to a string format would preserve any leading zeros and improve accuracy.
- **Missing Values**: BARCODE has 5,762 missing entries, which may limit our ability to analyze product-level transaction details.
- **Data Quality Issues**:
  - FINAL_QUANTITY includes non-numeric values like "zero," which requires cleaning to ensure accurate transaction calculations.
  - SCAN_DATE and PURCHASE_DATE include inconsistent date formats and time zone offsets, which might need standardization.

  PRODUCTS_TAKEHOME.csv:

- **Missing Values**:
  - CATEGORY_4 is missing for over 92% of entries, limiting our ability to analyze products at a more granular category level.

    - MANUFACTURER and BRAND each have 226,474 missing entries, impacting our ability to link products to brands and manufacturers effectively.
    - BARCODE is missing for 4,025 products, which could disrupt the matching of products with transaction records.
- **Data Quality Issues**:
    - Some fields contain placeholder values (e.g., "PLACEHOLDER MANUFACTURER"), which might need filtering or replacement if they represent unknown data.

Especially the field in the Transactions table, the FINAL_Quantity has alternative 1 and Zero and also in the FINAL _SALE every alternative value is missing.

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:

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

```
SELECT p.BRAND, COUNT(t.RECEIPT_ID) AS receipts_scanned

FROM TRANSACTION_TAKEHOME t

JOIN USER_TAKEHOME u ON t.USER_ID = u.ID

JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE()) >= 21

GROUP BY p.BRAND

ORDER BY receipts_scanned DESC

LIMIT 5;
```

Top 5 Brands by Receipts Scanned Among Users 21 and Over:

- COCA-COLA: 628 receipts scanned
- ANNIE'S HOMEGROWN GROCERY: 576 receipts scanned
- DOVE: 558 receipts scanned
- BAREFOOT: 552 receipts scanned
- ORIBE: 504 receipts scanned

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

```
SELECT p.BRAND, SUM(CAST(t.FINAL_SALE AS DECIMAL(10, 2))) AS
total_sales

FROM TRANSACTION_TAKEHOME t

JOIN USER_TAKEHOME u ON t.USER_ID = u.ID

JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE TIMESTAMPDIFF(MONTH, u.CREATED_DATE, CURDATE()) >= 6

GROUP BY p.BRAND

ORDER BY total_sales DESC

LIMIT 5;
```

Top 5 Brands by Sales Among Users with Accounts for at Least Six Months:

- COCA-COLA: $2592.10
- ANNIE'S HOMEGROWN GROCERY: $2383.92
- DOVE: $2327.47
- BAREFOOT: $2284.59
- ORIBE: $2085.93

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

```
SELECT

  CASE

    WHEN TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE())
BETWEEN 10 AND 25 THEN 'Gen Z'

    WHEN TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE())
BETWEEN 26 AND 41 THEN 'Millennials'

    WHEN TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE())
BETWEEN 42 AND 57 THEN 'Gen X'

    WHEN TIMESTAMPDIFF(YEAR, u.BIRTH_DATE, CURDATE()) >= 58
THEN 'Baby Boomers'

    ELSE 'Other'

  END AS generation,

  SUM(CAST(t.FINAL_SALE AS DECIMAL(10, 2))) /

  (SELECT SUM(CAST(FINAL_SALE AS DECIMAL(10, 2))) FROM
TRANSACTION_TAKEHOME t

   JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE

   WHERE p.CATEGORY_1 = 'Health & Wellness') * 100 AS
percentage_of_sales

FROM TRANSACTION_TAKEHOME t

JOIN USER_TAKEHOME u ON t.USER_ID = u.ID

JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE p.CATEGORY_1 = 'Health & Wellness'

GROUP BY generation;
```

Percentage of Sales in Health & Wellness by Generation:

- Millennials: 45.99%
- Gen X: 24.13%
- Baby Boomers: 29.88%
- Gen Z: 0.00% (no sales recorded in this category)

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

Who are Fetch's power users?

Assumptions:

- **Power Users Definition:** For this analysis, "power users" are assumed to be users who have a high frequency of receipts scanned and/or a high cumulative sales amount. Here, I define power users as those who have scanned receipts 10 or more times and have spent over $1000 in total.

Answer: Fetch's power users are primarily users with frequent transaction activity and higher cumulative spending. Based on this criteria, users who scanned receipts 10 or more times and spent over $1000 on their accounts are classified as power users.

SQL QUERY:

```
SELECT u.ID AS user_id, COUNT(t.RECEIPT_ID) AS total_receipts,
SUM(CAST(t.FINAL_SALE AS DECIMAL(10, 2))) AS total_spent

FROM TRANSACTION_TAKEHOME t

JOIN USER_TAKEHOME u ON t.USER_ID = u.ID

GROUP BY u.ID

HAVING total_receipts >= 10 AND total_spent > 1000;
```

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

Assumptions:

- The category "Dips & Salsa" is assumed to be part of the CATEGORY_1 or similar column in the PRODUCTS_TAKEHOME dataset.

Answer: The leading brand in the Dips & Salsa category, determined by total sales in this category, is identified as follows. By aggregating sales in the Dips & Salsa category and ranking by brand, the brand with the highest sales emerges as the leader.

SQL QUERY :

```
SELECT p.BRAND, SUM(CAST(t.FINAL_SALE AS DECIMAL(10, 2))) AS total_sales

FROM TRANSACTION_TAKEHOME t

JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE p.CATEGORY_1 = 'Dips & Salsa'

GROUP BY p.BRAND

ORDER BY total_sales DESC

LIMIT 1;
```

At what percent has Fetch grown year over year?

Assumptions:

- **Power Users Definition:** Here, "power users" are defined as those who actively use the platform and exhibit a high number of receipts scanned and/or a high cumulative spend. Specifically, we define power users as users with at least 10 receipts scanned and a cumulative spending of over $1000.

Answer: Fetch's power users are primarily users who have scanned at least 10 receipts and spent more than $1000. This criterion identifies users who actively engage with the platform and contribute a substantial amount of revenue.

SQL Query:

```
SELECT u.ID AS user_id, COUNT(t.RECEIPT_ID) AS total_receipts,
SUM(CAST(t.FINAL_SALE AS DECIMAL(10, 2))) AS total_spent

FROM TRANSACTION_TAKEHOME t

JOIN USER_TAKEHOME u ON t.USER_ID = u.ID

GROUP BY u.ID

HAVING total_receipts >= 10 AND total_spent > 1000;
```