FETCH Take Home Assessment

Mohammed Jawhar

Review the unstructured csv files and answer the following questions with code that supports your conclusions:

- **Are there any data quality issues present?**

For Users: (100000 records)

ID: Zero NULL values, 100000 unique values, ID values are all Unique

CREATED DATE:

Has zero NULL values, 99942 unique values. Accounts starting from 2014-04-18 to 2024-09-11. Most account creation in 2022, 2023 and 2021 in that order.

BIRTH DATE:

3675 NULL values, 54721 unique values.

Birth dates starting from 1900-01-01 to 2022-04-03. Most frequent birth dates between 1980-2005

STATE: 4812 NULL Records, Total Unique 52 records - 50 US States, Nan, DC(US territory). Most User state is TX and the state with the least users is VT. Top 5 states i.e., TX, FL, CA, NY, IL - makes up 37.85% of the users.

LANGUAGE: 30508 NULL records. Total Unique 2 records - Eng and es-419. 91.2% eng speakers and 8.7% Spanish speakers

GENDER: 5892 NULL Records. 7 unique records - female, male, non_binary, transgender, prefer_not_to_say, unknown, not_listed. Most frequent gender is female with 68.39% and least frequent is not_listed with 0.029%.

For Products: (845552 records)

CATEGORY_1 - 27 unique categories (including NaN). NULL values : 0.013%. Health & Wellness makes up 60.64% of the entries and Snacks makes up 38.41% of the entries. 99% of the entries are "Health & Wellness" and "Snacks".

CATEGORY_2 - 121 unique categories (including NaN). NULL values : 0.16%. Top 5 makeup 56.35% of the NON NULL categories and top 10 categories make up 72.36% of the NON NULL values.

CATEGORY_3 - 344 unique categories (including NaN). NULL values : 7.16%. Top 5 categories: Confection Candy, Vitamins & Herbal Supplements, Chocolate Candy, Hair Styling Products, Reading Glasses.

CATEGORY_4 - 127 unique categories (including NaN). NULL values : 92.02%. The top 10 categories make up 64% of the NON NULL values.

MANUFACTURER - 4354 unique manufacturers (including NaN). NULL values : 26.78%.

BRAND - 8122 unique brands (including NaN). NULL values : 26.78%

BARCODE - NULL values : 0.47%. 841342 unique barcodes

For Transactions: (50000 records)

RECEIPT_ID - Zero NULL Records. 24440 unique RECEIPT IDs and 25560 duplicates.

PURCHASE_DATE - Zero NULL Records. 89 unique Purchase Dates. Purchase dates between 2024-06-12 and 2024-09-08.

SCAN_DATE - Zero NULL Records. 24440 unique Scanning Dates. Purchase dates between 2024-06-12 and 2024-09-08.

STORE_NAME - Zero NULL Records. 42.6% of the records are from WALMART.

USER_ID - Zero NULL Records. 17694 unique User Ids.

BARCODE - 11.5% NULL Records. 11027 unique barcodes.

FINAL_QUANTITY - 12500 NULL Records. 25% NULL records. 86 unique records. 276 is the max final_quantity.

FINAL_SALE - 12500 NULL records. 25% NULL records. 1434 unique records. 462.8 is the max final_quantity.

DATA JOINS:

- There are 834960 products without any transactions
- There are 19,408 transactions without any product information.
- There are barcodes under 5 characters in length where the standard BARCODE value in the table is at least 14 characters length.
- There are user_ids present in transactions but not in the users table

Are there any fields that are challenging to understand?

There needs to be clarity in the following:

- Es-419 in users.LANGUAGE - does it include any Spanish dialect or only the ones spoken in Latin America and Caribbean.
- transactions.FINAL_QUANTITY - what does it represent?
- transactions.FINAL_SALE - what does it represent? Do we need to multiply it with FINAL_QUANTITY to get the total sale value?

Closed-ended questions:

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

```
q1 = """ WITH Users21Plus AS (
    SELECT ID
    FROM users
    WHERE BIRTH_DATE <= strftime('%Y-%m-%d', 'now', '-21 years')
),
FilteredTransactions AS (
    SELECT t.RECEIPT_ID, t.BARCODE
    FROM transactions t
    JOIN Users21Plus u ON t.USER_ID = u.ID
)
SELECT p.BRAND, COUNT(DISTINCT f.RECEIPT_ID) AS reciept_count
FROM FilteredTransactions f
JOIN products p ON f.BARCODE = p.BARCODE
WHERE p.BRAND IS NOT NULL
GROUP BY p.BRAND
ORDER BY reciept_count DESC
LIMIT 5
"""
print(ps.sqldf(q1, locals()))
```

```
              BRAND  reciept_count
0      NERDS CANDY              3
1             DOVE              3
2          TRIDENT              2
3  SOUR PATCH KIDS              2
4           MEIJER              2
```

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

```
q2 = """ WITH Users6Months AS (
    SELECT ID
    FROM users
    WHERE CREATED_DATE <= strftime('%Y-%m-%d', 'now', '-6 months')
),
FilteredTransactions AS (
    SELECT t.USER_ID, t.BARCODE, CAST(t.FINAL_SALE AS FLOAT) AS SALE_AMOUNT
    FROM transactions t
    JOIN Users6Months u ON t.USER_ID = u.ID
    WHERE t.FINAL_SALE IS NOT NULL AND t.FINAL_SALE != ''
)
SELECT p.BRAND, SUM(f.SALE_AMOUNT) AS total_sales
FROM FilteredTransactions f
JOIN products p ON f.BARCODE = p.BARCODE
WHERE p.BRAND IS NOT NULL
GROUP BY p.BRAND
ORDER BY total_sales DESC
LIMIT 5;
"""
print(ps.sqldf(q2, locals()))
```

```
          BRAND  total_sales
0           CVS        72.00
1      TRIDENT        46.72
2         DOVE        42.88
3  COORS LIGHT        34.96
4       QUAKER        16.60
```

Open-ended questions:

Who are Fetch's power users?
Assumption: Power users are defined as users with the highest number of transactions and the highest total spend.

```
: oq1 ="""SELECT
      t.USER_ID,
      COUNT(t.RECEIPT_ID) AS total_transactions,
      SUM(t.FINAL_SALE) AS total_spent
  FROM transactions t
  WHERE t.FINAL_SALE IS NOT NULL
  GROUP BY t.USER_ID
  ORDER BY total_transactions DESC, total_spent DESC
  LIMIT 10;"""
  print(ps.sqldf(oq1, locals()))
```

	USER_ID	total_transactions	total_spent
0	64e62de5ca929250373e6cf5	22	57.65
1	62925c1be942f00613f7365e	20	49.87
2	604278958fe03212b47e657b	20	46.61
3	64063c8880552327897186a5	18	43.72
4	60a5363facc00d347abadc8e	14	101.97
5	609af341659cf474018831fb	14	25.55
6	61d5f5d2c4525a3a478b386b	14	25.22
7	624dca0770c07012cd5e6c03	14	21.91
8	6327a07aca87b39d76e03864	14	15.64
9	653a0f40909604bae9071473	12	84.78

Reasoning for Identifying Fetch's Power Users
When defining power users, it is crucial to go beyond just total spending as a metric. While some users may have high total spend values, they might have only made one or two transactions, indicating that they are one-time shoppers rather than engaged, repeat customers.

Why Not Use Only Total Spend?
Users with high total spend but only 1-2 transactions are not repeat customers.
Loyalty and retention are more valuable than one-time high-value purchases.
A high transaction count combined with consistent spending is a stronger indicator of customer engagement and loyalty.

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: Key Insights & Data Quality Concerns from Fetch Data Analysis
Dear Product Leader,

The initial analysis of the Fetch data has concluded, and I wanted to share key findings, data quality concerns, and some areas where we need additional clarification.

I. **Key Data Quality Issues**
Products Without Transactions – Over 834,960 products in the dataset have never been purchased, which could indicate inactive SKUs or data ingestion gaps.
Missing Product Data in Transactions – 19,408 transactions lack associated product details, making it unclear what was purchased.
Users in Transactions But Not in Users Table – Some transactions are linked to USER_IDs that do not exist in the Users dataset, suggesting missing records or integration issues.
Barcodes in Scientific Notation – Discrepancies in BARCODE information especially of BARCODES with length<5.

II. **Interesting Trend in the Data**
We investigated power users and found that total spend alone is not a reliable metric. Some of the highest-spending users only made one or two transactions, meaning they are not repeat customers. Instead, power users should be defined by both high transaction frequency and consistent spending, ensuring they are loyal and engaged customers.

III. **Request for Action**
To resolve the outstanding data issues and refine our insights, we need:
Clarification on Missing Product Data – Are transactions missing product info due to scanning errors, or are certain products excluded from reporting?
User Data Validation – Can we confirm whether all USER_IDs should exist in the Users table? If not, how do we handle unmatched users?
Barcode Format Fixes – Should we reformat BARCODE to ensure proper linking between products and transactions?

Let me know if we can set up a quick sync to address these issues and improve data reliability for future analysis.

Best Regards, Jawhar M