# Fetch - Data Analyst Take Home

**Rui Pan**

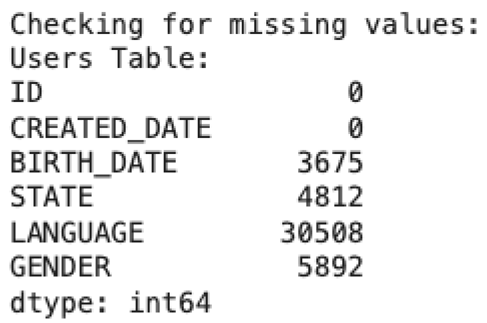
## First: explore the data

### 1. Are there any data quality issues present?

Yes, there are several data quality issues in the dataset:

### 1)Missing Values

**Users Table**



```
Checking for missing values:
Users Table:
ID                 0
CREATED_DATE       0
BIRTH_DATE      3675
STATE           4812
LANGUAGE       30508
GENDER          5892
dtype: int64
```

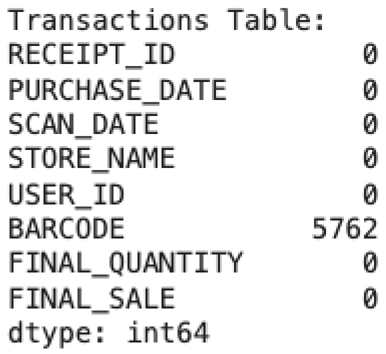
**BIRTH_DATE**: 3,675 missing values, which will impact any age-based analyses, such as filtering users by age (e.g., users 21 years and older).

**STATE**: 4,812 missing values, which could affect geographic segmentation or location-based analyses.

**LANGUAGE**: 30,508 missing values, which is significant and will limit any analysis or segmentation based on user language preferences.

**GENDER**: 5,892 missing values, which will affect gender-based analysis and segmentation.

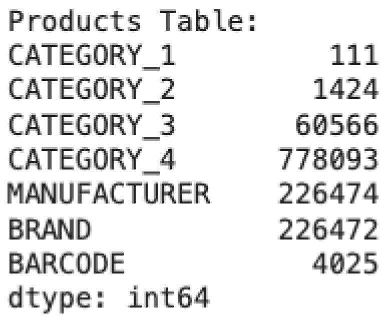
### Transactions Table:



```
Transactions Table:
RECEIPT_ID           0
PURCHASE_DATE        0
SCAN_DATE            0
STORE_NAME           0
USER_ID              0
BARCODE           5762
FINAL_QUANTITY       0
FINAL_SALE           0
dtype: int64
```

**BARCODE**: 5,762 missing values, which are critical because they prevent linking transactions to specific products. Any analysis involving products (e.g., brand performance, category sales) will be incomplete for these transactions.

**Products Table:**



```
Products Table:
CATEGORY_1          111
CATEGORY_2         1424
CATEGORY_3        60566
CATEGORY_4       778093
MANUFACTURER     226474
BRAND            226472
BARCODE            4025
dtype: int64
```

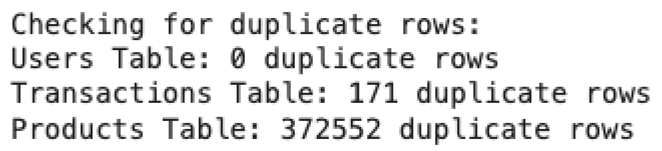
**CATEGORY_1**: 111 missing values, which is important because this is the highest-level product categorization. Missing values here will impact category-level analyses.

**CATEGORY_2, CATEGORY_3, CATEGORY_4**: The number of missing values increases with more granular product categories, with **CATEGORY_4** having 778,093 missing values. These could limit the depth of analysis for more detailed product hierarchies.

**MANUFACTURER and BRAND**: Approximately 226,000 missing values for each. This will affect any brand or manufacturer-specific analysis.

**BARCODE**: 4,025 missing values, which will cause issues linking products to transactions, impacting sales and product-level analyses.

**2) Duplicate Data Issues**



```
Checking for duplicate rows:
Users Table: 0 duplicate rows
Transactions Table: 171 duplicate rows
Products Table: 372552 duplicate rows
```

### Transactions Table:

There are 171 duplicate rows, which means some transactions are recorded more than once. This could cause problems when analyzing purchase behavior or transaction totals. These duplicates need to be removed to ensure the data is correct and reliable.

### **Products Table**:

There are 372,552 duplicate rows, which is a major issue. Having so many duplicate products can distort the analysis, especially for things like tracking brand performance or sales by category. It's important to clean these up to prevent misleading results.

## 2. Are there any fields that are challenging to understand?

**GENDER**: It's unclear if the language field means the language the user prefers for the app or the user's native language. Knowing this would help us use the data better when grouping users.

**LANGUAGE**: These seem like deeper levels of product categories, but many are missing. It's important to know how these categories connect to each other so we can better handle missing data and understand results.

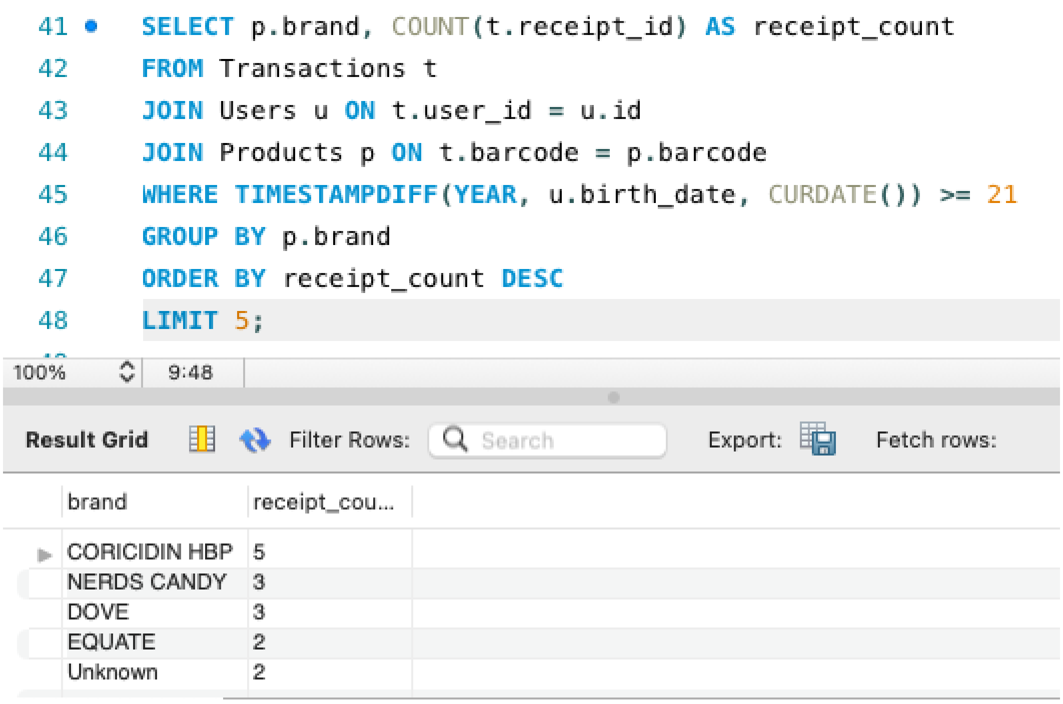
**CATEGORY_2, CATEGORY_3, CATEGORY_4**: The hierarchical structure of these categories might not be immediately clear, especially since there are significant numbers of missing values in deeper categories. Understanding how these categories relate to each other would help in making assumptions about missing data or in interpreting results.

**BARCODE**: We need to understand if barcodes are specific to products, stores, or something else. This would help us deal with missing or duplicate barcodes more effectively.

# Second: provide queries

### Closed-ended questions:

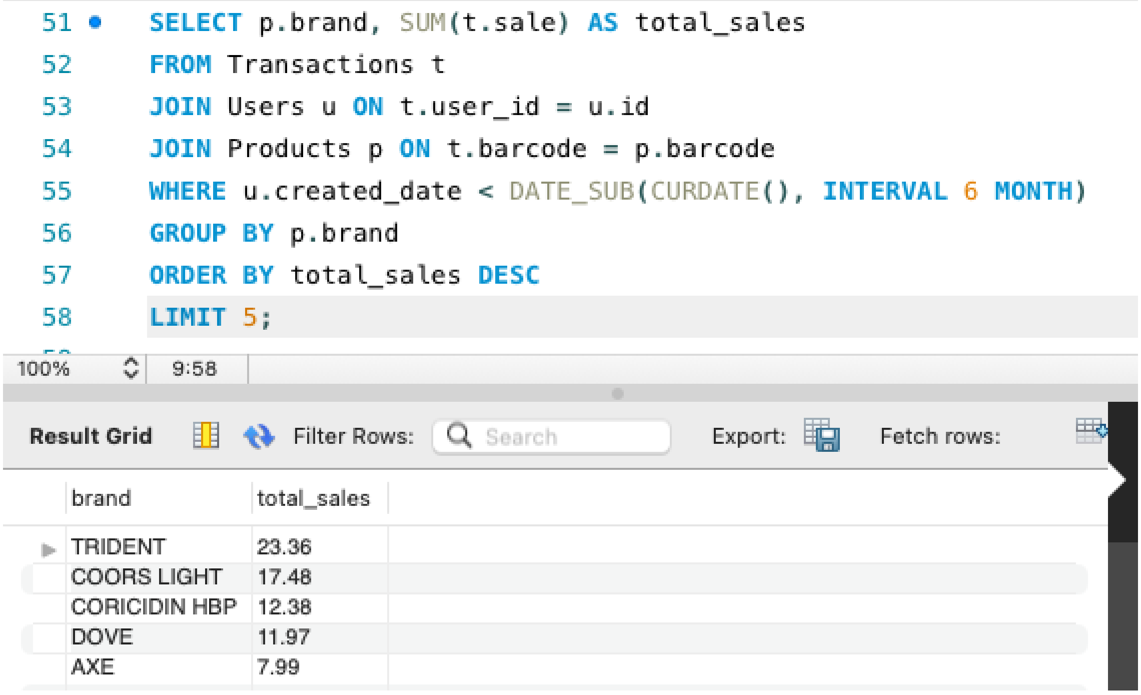
### 1) What are the top 5 brands by receipts scanned among users 21 and over?



```
SELECT p.brand, COUNT(t.receipt_id) AS receipt_count
FROM Transactions t
JOIN Users u ON t.user_id = u.id
JOIN Products p ON t.barcode = p.barcode
WHERE TIMESTAMPDIFF(YEAR, u.birth_date, CURDATE()) >= 21
GROUP BY p.brand
ORDER BY receipt_count DESC
LIMIT 5;
```

```
brand           receipt_cou...
CORICIDIN HBP   5
NERDS CANDY     3
DOVE            3
EQUATE          2
Unknown         2
```

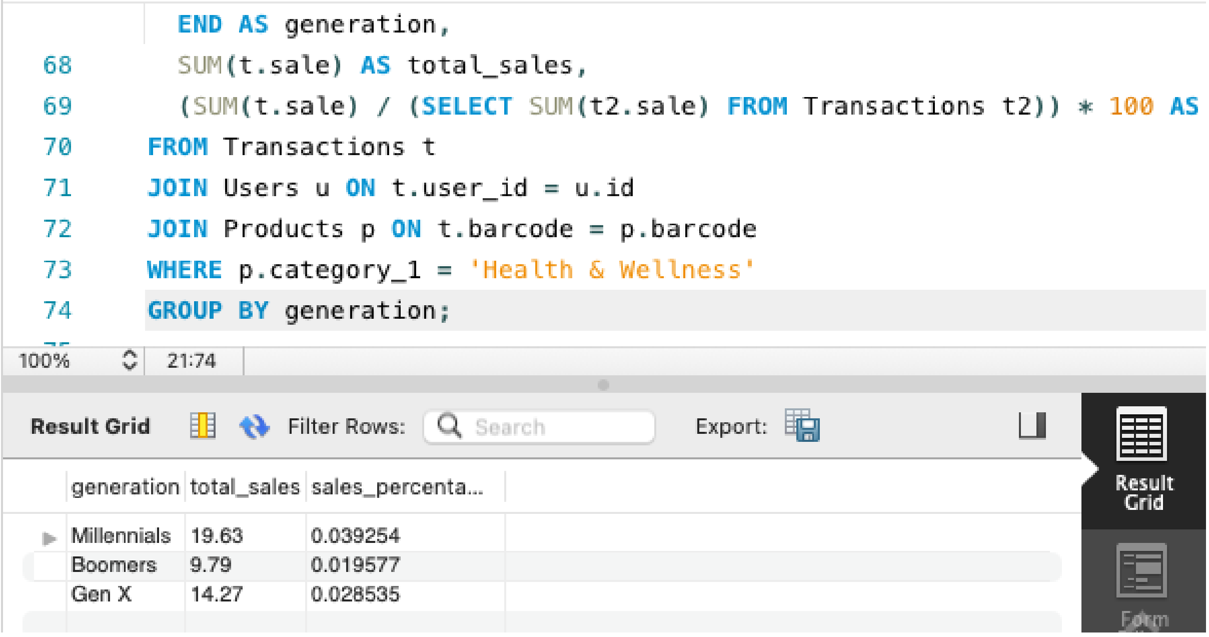
2）What are the top 5 brands by sales among users that have had their account for at least six months?



```
SELECT p.brand, SUM(t.sale) AS total_sales
FROM Transactions t
JOIN Users u ON t.user_id = u.id
JOIN Products p ON t.barcode = p.barcode
WHERE u.created_date < DATE_SUB(CURDATE(), INTERVAL 6 MONTH)
GROUP BY p.brand
ORDER BY total_sales DESC
LIMIT 5;
```

```
brand           total_sales
TRIDENT         23.36
COORS LIGHT     17.48
CORICIDIN HBP   12.38
DOVE            11.97
AXE             7.99
```

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

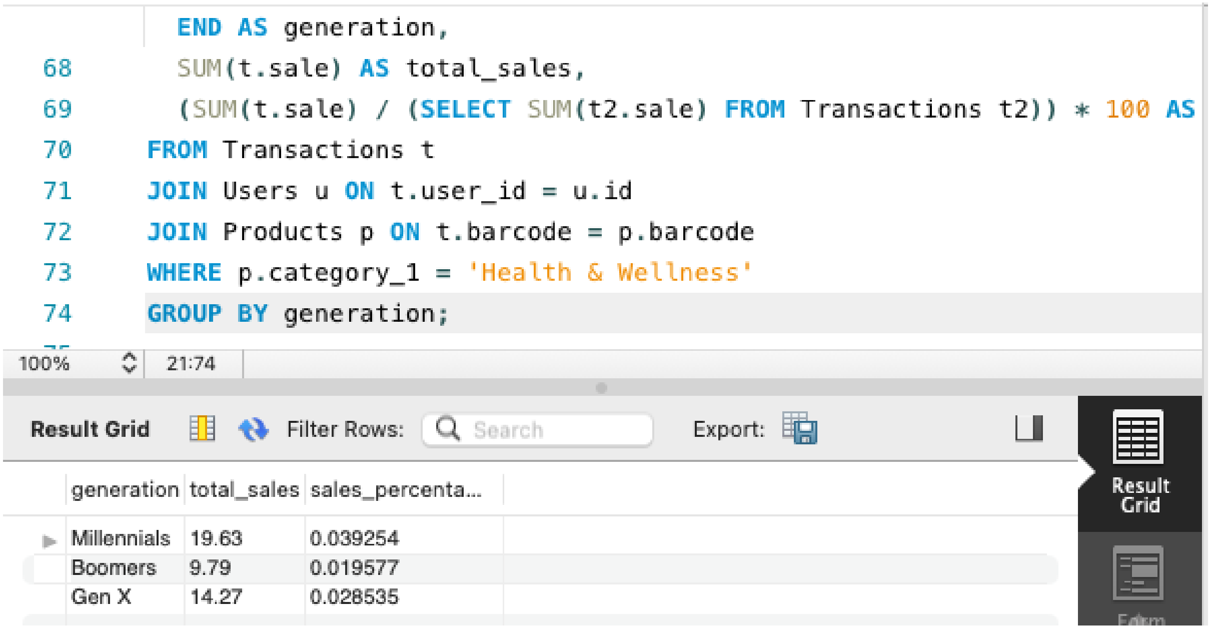


```
      END AS generation,
      SUM(t.sale) AS total_sales,
      (SUM(t.sale) / (SELECT SUM(t2.sale) FROM Transactions t2)) * 100 AS
    FROM Transactions t
    JOIN Users u ON t.user_id = u.id
    JOIN Products p ON t.barcode = p.barcode
    WHERE p.category_1 = 'Health & Wellness'
    GROUP BY generation;
```

| generation | total_sales | sales_percenta... |
|---|---|---|
| Millennials | 19.63 | 0.039254 |
| Boomers | 9.79 | 0.019577 |
| Gen X | 14.27 | 0.028535 |

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

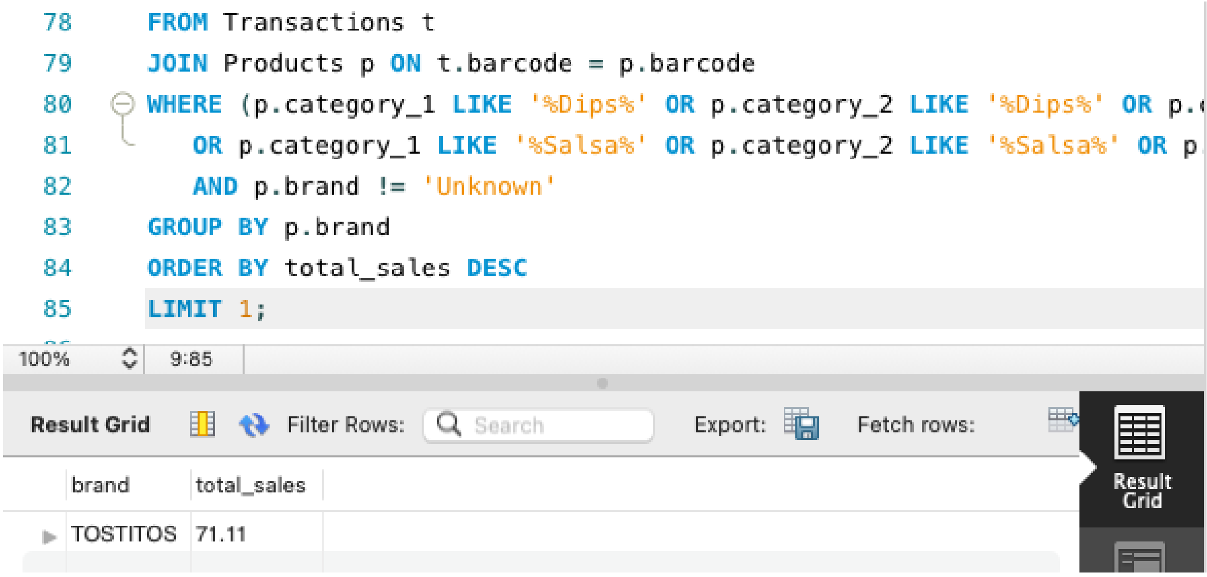
### 1) Who are Fetch's power users?



```
      END AS generation,
      SUM(t.sale) AS total_sales,
      (SUM(t.sale) / (SELECT SUM(t2.sale) FROM Transactions t2)) * 100 AS
    FROM Transactions t
    JOIN Users u ON t.user_id = u.id
    JOIN Products p ON t.barcode = p.barcode
    WHERE p.category_1 = 'Health & Wellness'
    GROUP BY generation;
```

| generation | total_sales | sales_percenta... |
|---|---|---|
| Millennials | 19.63 | 0.039254 |
| Boomers | 9.79 | 0.019577 |
| Gen X | 14.27 | 0.028535 |

### 2) Which is the leading brand in the Dips & Salsa category?

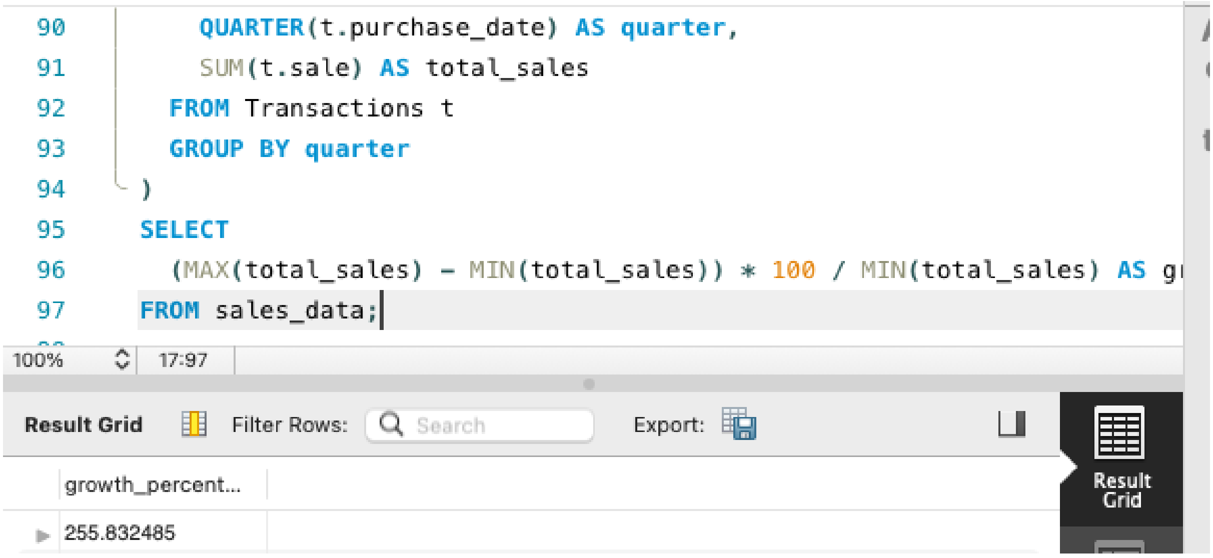


```sql
FROM Transactions t
JOIN Products p ON t.barcode = p.barcode
WHERE (p.category_1 LIKE '%Dips%' OR p.category_2 LIKE '%Dips%' OR p.
   OR p.category_1 LIKE '%Salsa%' OR p.category_2 LIKE '%Salsa%' OR p
   AND p.brand != 'Unknown'
GROUP BY p.brand
ORDER BY total_sales DESC
LIMIT 1;
```

```
brand      total_sales
TOSTITOS   71.11
```

While analyzing the Dips & Salsa category, I found that many products in the Products table had the brand field marked as "Unknown," indicating incomplete data. To ensure accurate results, I filtered out these "Unknown" brands, focusing only on products with valid brand names. This led to the identification of Tostitos as the leading brand in the category, with 71.11 in total sales. By excluding incomplete data, the result more accurately reflects user preferences.

### 3) At what percent has Fetch grown year over year?



```sql
    QUARTER(t.purchase_date) AS quarter,
    SUM(t.sale) AS total_sales
  FROM Transactions t
  GROUP BY quarter
)
SELECT
  (MAX(total_sales) - MIN(total_sales)) * 100 / MIN(total_sales) AS g
FROM sales_data;
```

```
growth_percent...
255.832485
```

I calculated Fetch's growth using quarterly sales data instead of yearly data because the available data was only for a limited period of 2024. Based on that, Fetch has grown by about 255.83% quarter-over-quarter. This is a very high growth rate, but I want to point out that using quarterly data instead of yearly might make the figure seem larger than if we had full yearly data for comparison.

## Third: communicate with stakeholders

Hi Stakeholders,

I wanted to share the results from my analysis of Fetch’s user and transaction data. Below is a summary of the key insights, some concerns about data quality, and what steps we might take next to dig deeper.

**Key Findings:**

**1.Top Brands by Sales**:

For users who have been with Fetch for over six months, Trident is the leading brand by sales, followed by Coors Light and Coricidin HBP. This suggests that well-known brands across different categories continue to perform strongly over time.

Millennials have the highest sales in the Health & Wellness category, contributing around 39% of total sales in that category, followed by Gen X and Boomers. This shows that younger generations are more engaged in purchasing health-related products through Fetch.

Tostitos is the top-performing brand in the Dips & Salsa category with significant sales, suggesting that users frequently purchase this brand when shopping for snacks.

**2. Data Quality Concerns:**

There are several missing fields in key tables:

Fields like language and state have a large number of missing entries.

Over 372,000 duplicate entries were found. Also, many brand and manufacturer fields are either missing or marked as "Unknown," which affects our ability to identify leading brands accurately.

A few duplicate entries were identified that could distort insights.

**3. Trend Observed:**

One interesting trend I noticed is the strong performance of well-known consumer brands like Trident and Tostitos, which consistently appear at the top in their respective categories. This suggests that Fetch users are more likely to engage with familiar brands when redeeming rewards, which is something we could leverage in future marketing campaigns.

**4. Next Steps / Request for Action:**

To better understand user behavior, it would help to fill in the missing user profile information (like language and state) and clean up the product information (remove duplicates and complete missing brands).

I'd recommend setting up validation processes to prevent missing data fields and duplicates going forward.

I’d like to further explore why some brands are marked as "Unknown" in the system and whether this is due to poor data entry or system errors.

Could we schedule a short discussion to address these points and clarify the next steps?

Best regards,
Rui Pan