1. Data Quality Issues and Cleaning

Issues Identified

1. **Missing Data**:
    - BARCODE (Transactions): 11.52% missing (5,762 missing out of 50,000 records).
    - **CATEGORY_4, MANUFACTURER, BRAND** (Products): Significant missing values, ***CATEGORY_4*** is ***92.02%*** missing, so dropping category 4
    - There isn't proper hierarchy between category 1, category 2, category 3.
    - BIRTH_DATE (Users): 3.68% missing (3,675 missing out of 100,000 records).
2. **Inconsistent Data Types**:
    - FINAL_QUANTITY (Transactions) contains non-numeric values such as "zero".
    - Dates like CREATED_DATE and PURCHASE_DATE are stored as strings instead of datetime.
    - Primary key, Foreign keys like BARCODE have so much NaN values which lacks data integrity.
3. **Ambiguous Values**:
    - FINAL_QUANTITY: It is unclear if "zero" indicates an actual quantity of 0 or an error.

IsNULL Values: % of Null values

	0
CATEGORY_1	0.013128
CATEGORY_2	0.168411
CATEGORY_3	7.162895
CATEGORY_4	92.021898
MANUFACTURER	26.784160
BRAND	26.783923
BARCODE	0.476020

Cleaning Steps

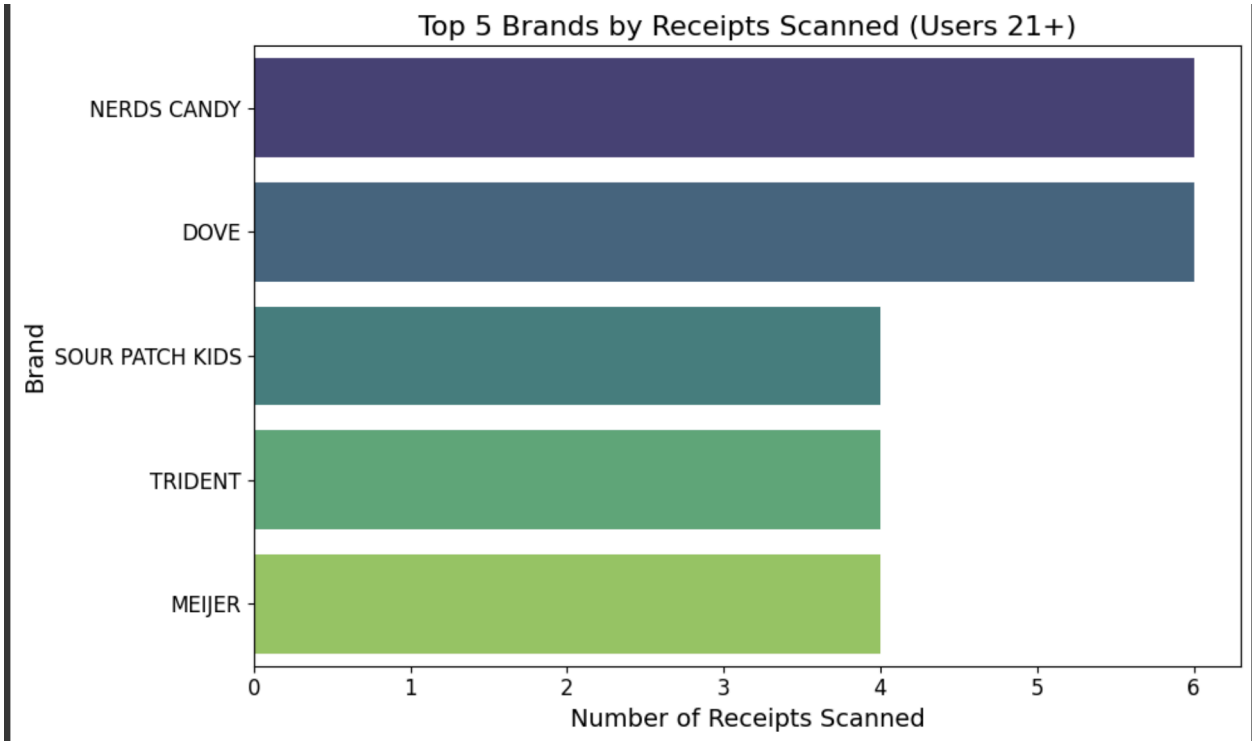
1. Converted date columns (CREATED_DATE, PURCHASE_DATE, SCAN_DATE) to datetime format.
2. Replaced ambiguous values in FINAL_QUANTITY:
    - Non-numeric values (e.g., "zero") converted to 0.
3. Filled missing values:

- Replaced missing CATEGORY_1 with "Others".
- Replaced missing CATEGORY_2 and CATEGORY_3 with values like "Other Category".
- Replaced missing BRAND and MANUFACTURER with "Unknown".

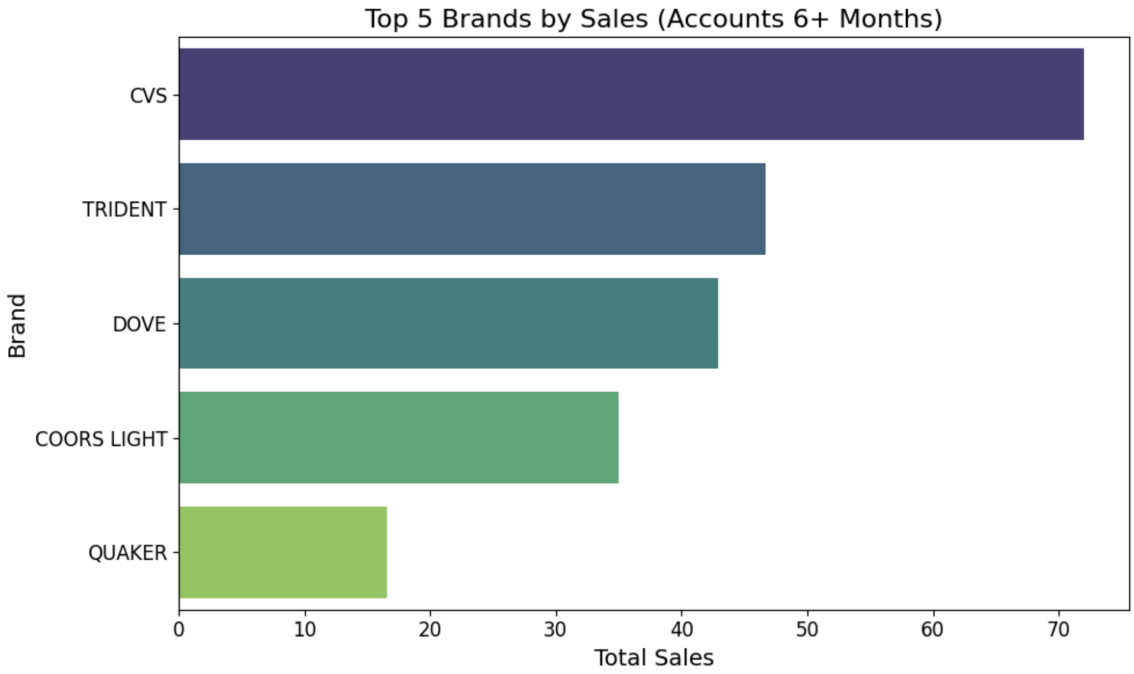
4. Dropped rows with missing BARCODE in the TRANSACTIONS table. # but have named it 000 for general analysis purchase

ANALYSIS:

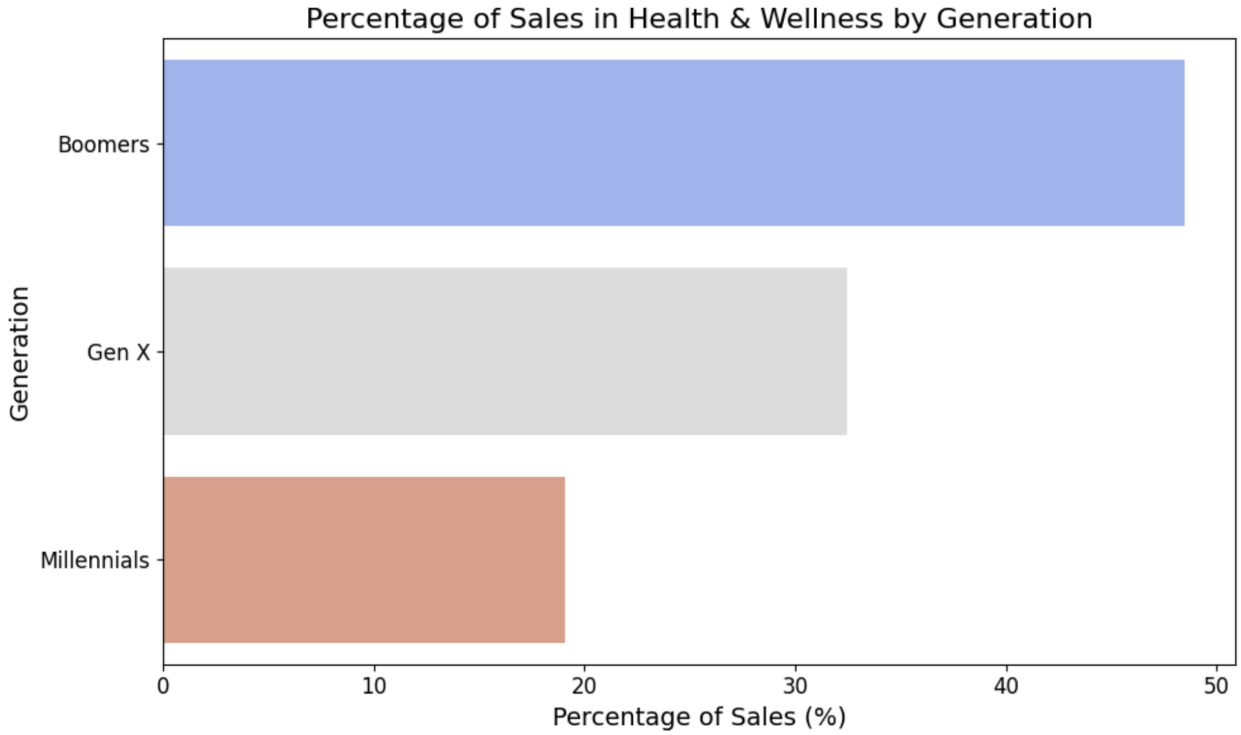
Top 5 Brands by Receipts Scanned Among Users 21 and Over?



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



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



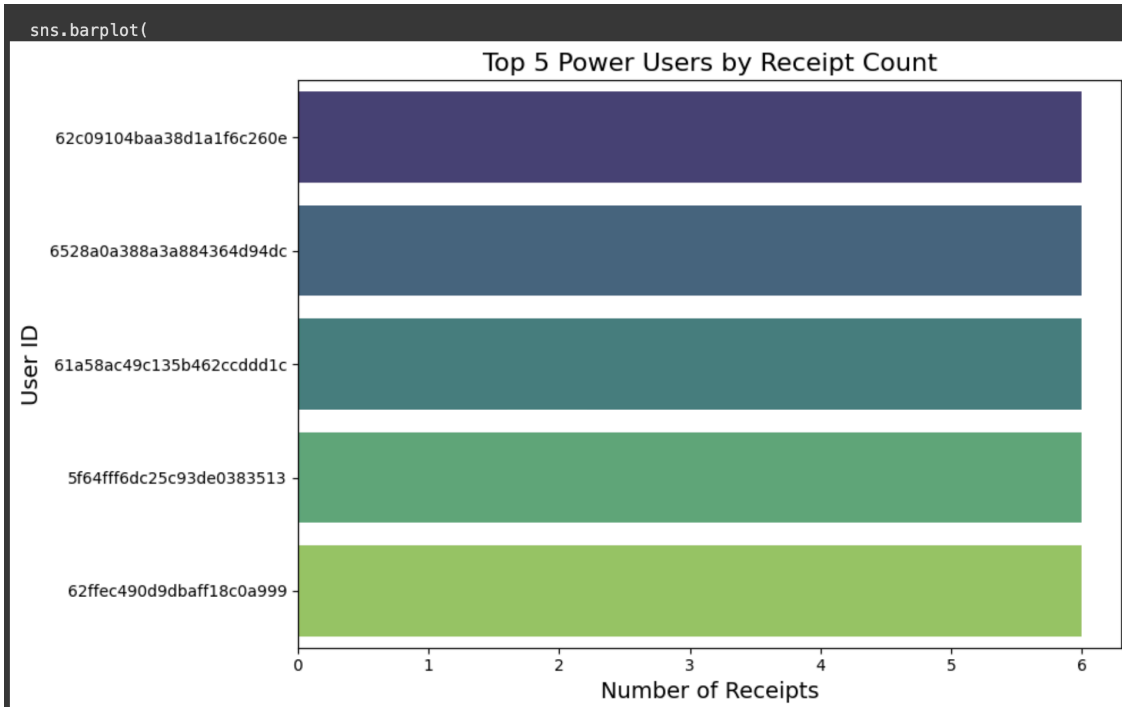
1. Who Are Fetch's Power Users?

Assumptions:

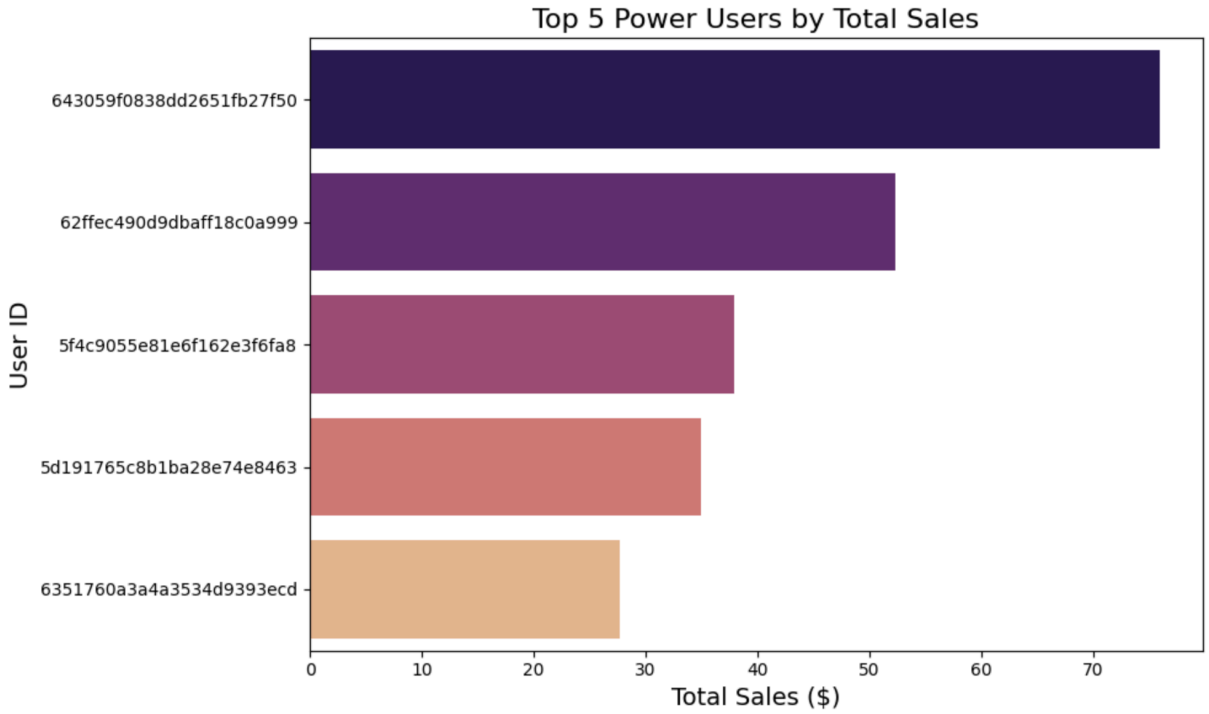
1. Power users are defined based on two metrics:
    - **Engagement**: Users with the highest number of distinct receipts scanned (**`RECEIPT_ID`**).
    - **Spending**: Users with the highest total sales (**`FINAL_SALE`**).
2. Users must have participated in transactions during the observation period to qualify as power users.

Answer:

- **Engagement**: The top 5 power users by receipts scanned are those who have scanned the most unique receipts.



- **Spending**: The top 5 power users by spending are users who contributed the most to total sales.



Recommendation: Target these users with loyalty programs or exclusive offers to encourage continued engagement and higher spending.

2. Which Is the Leading Brand in the Dips & Salsa Category?

Assumptions:

1. The "leading brand" is determined by the total sales (`FINAL_SALE`) in the ***`Dips & Salsa`*** category (***`CATEGORY_2`***).
2. Sales data from transactions accurately reflects brand performance.
3. Missing or null values in the **`BRAND`** column are excluded from the analysis.

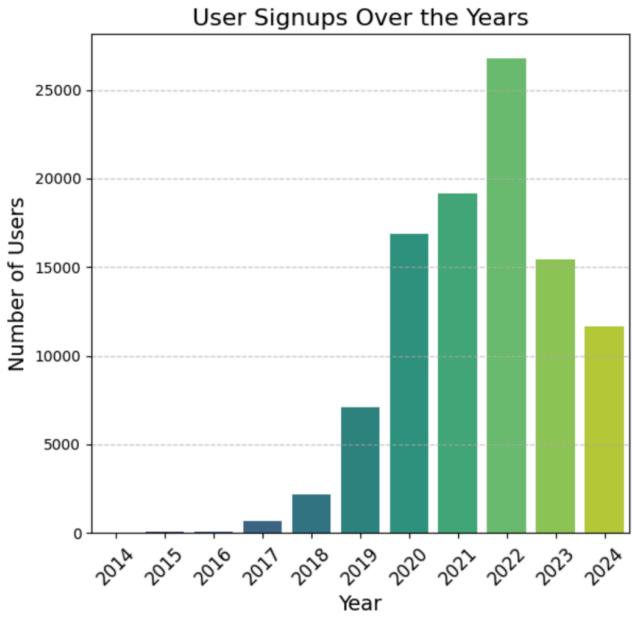
Answer: ***Leading Brand in Dips & Salsa Category: MARZETTI with sales of $ 5.99***

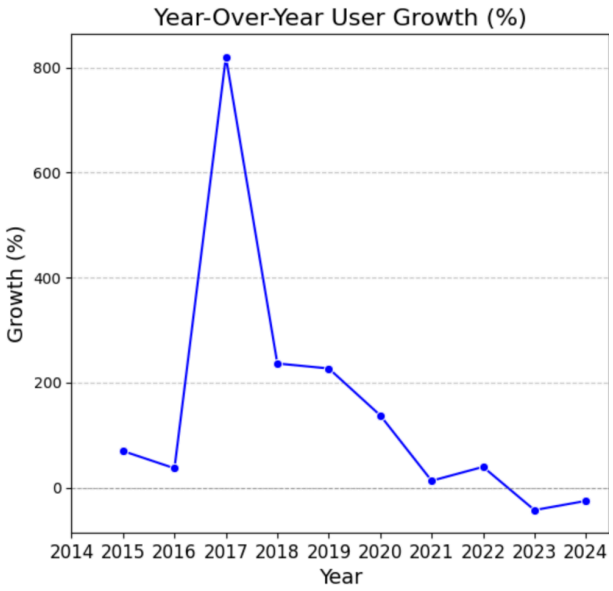
3. At What Percent Has Fetch Grown Year Over Year?

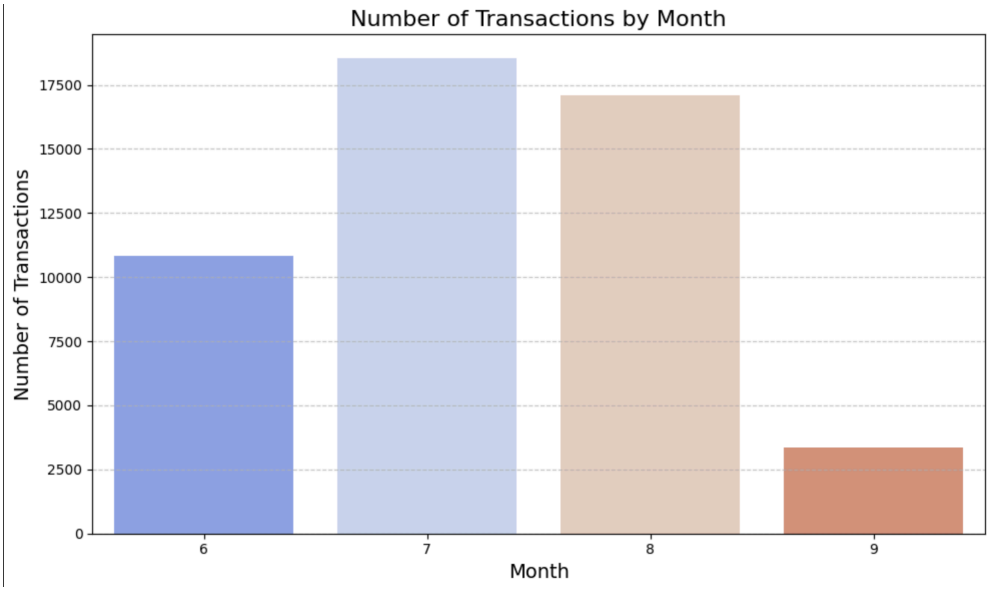
Assumptions:

- **Growth Metric**: Based on user signups using **`CREATED_DATE`** (as purchase data is only available for the last 6 months).
- **Growth Rate**: Calculated as the year-over-year percentage change in user registrations.

Trends:







1. **Rapid Growth**: Fetch experienced its highest growth in 2017 (820%), likely due to increased user acquisition strategies or awareness.
2. **Stabilization**: Growth slowed significantly after 2020, indicating a shift to user retention strategies.
3. **Decline**: User growth dropped in 2023 and 2024, requiring investigation into the factors causing the slowdown.

Recommendation: Focus on addressing the decline in user growth by:

- Revisiting marketing strategies.
- Engaging inactive users through loyalty programs.
- Focus on retention strategies for newly acquired users to maximize the impact of recent growth.
- Use insights from high-growth periods to replicate successful marketing initiatives.

SQL QUERIES

Top 5 Brands by Receipts Scanned Among Users 21 and Over

```
SELECT
   p.BRAND,
   COUNT(DISTINCT t.RECEIPT_ID) AS receipt_count
FROM
   USERS u
JOIN
   TRANSACTIONS t ON u.ID = t.USER_ID
JOIN
   PRODUCTS p ON t.BARCODE = p.BARCODE
WHERE
   (YEAR(CURDATE()) - YEAR(u.BIRTH_DATE)) >= 21
GROUP BY
   p.BRAND
ORDER BY
   receipt_count DESC
LIMIT 5;
```

Top 5 Brands by Sales Among Users with Accounts 6+ Months

```
SELECT
   p.BRAND,
   SUM(t.FINAL_SALE) AS total_sales
FROM
   USERS u
JOIN
   TRANSACTIONS t ON u.ID = t.USER_ID
JOIN
   PRODUCTS 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;
```

Percentage of Sales in Health & Wellness Category by Generation

```
WITH GENERATION_SALES AS (
    SELECT
        CASE
            WHEN YEAR(BIRTH_DATE) <= 1945 THEN 'Silent'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1946 AND 1964 THEN 'Boomers'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1965 AND 1980 THEN 'Gen X'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1981 AND 1996 THEN 'Millennials'
            WHEN YEAR(BIRTH_DATE) >= 1997 THEN 'Gen Z'
            ELSE 'Unknown'
        END AS generation,
        SUM(t.FINAL_SALE) AS health_sales
    FROM
        USERS u
    JOIN
        TRANSACTIONS t ON u.ID = t.USER_ID
    JOIN
        PRODUCTS p ON t.BARCODE = p.BARCODE
    WHERE
        p.CATEGORY_1 = 'Health & Wellness'
    GROUP BY
        generation
),
TOTAL_SALES AS (
    SELECT
        CASE
            WHEN YEAR(BIRTH_DATE) <= 1945 THEN 'Silent'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1946 AND 1964 THEN 'Boomers'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1965 AND 1980 THEN 'Gen X'
            WHEN YEAR(BIRTH_DATE) BETWEEN 1981 AND 1996 THEN 'Millennials'
            WHEN YEAR(BIRTH_DATE) >= 1997 THEN 'Gen Z'
            ELSE 'Unknown'
        END AS generation,
        SUM(t.FINAL_SALE) AS total_sales
    FROM
        USERS u
    JOIN
        TRANSACTIONS t ON u.ID = t.USER_ID
```

```
    GROUP BY
        generation
)
SELECT
    g.generation,
    (g.health_sales / t.total_sales) * 100 AS percentage_of_sales
FROM
    GENERATION_SALES g
JOIN
    TOTAL_SALES t ON g.generation = t.generation;
```