Fetch Data Analysis Report

1. Introduction

This report analyzes the provided datasets for Fetch, focusing on data quality assessment, SQL query-based insights, and stakeholder communication. The objective is to evaluate data integrity, extract key business insights, and summarize the findings in a way that is useful for decision-making.

2. Data Exploration and Quality Assessment

2.1 Overview of Datasets

The analysis was conducted on three datasets:

- **USER_TAKEHOME.csv**: Contains user demographic information.
- **TRANSACTION_TAKEHOME.csv**: Records transaction details.
- **PRODUCTS_TAKEHOME.csv**: Lists product details including categories and brands.

2.2 Identified Data Quality Issues

- **Missing Values:**
  - `USER`: `BIRTH_DATE` (3.67%), `STATE` (4.81%), `LANGUAGE` (30.51%), `GENDER` (5.89%).
  - `TRANSACTION`: `BARCODE` (11.52%).
  - `PRODUCTS`: `CATEGORY_4` (92.02%), `MANUFACTURER` (26.78%), `BRAND` (26.78%).
- **Duplicate Records:**
  - `TRANSACTION`: 171 duplicates.
  - `PRODUCTS.csv`: 215 duplicates.
- **Inconsistencies in `FINAL_QUANTITY` and `FINAL_SALE` values:**
  - Entries containing 'zero' and blank spaces instead of nulls.
  - Repeated transactions with different quantities and sales values.

2.3 Further Investigation on Transactions data

- Investigated different combinations of quantity and sales value within the same receipt and barcode. More than 50% of records exhibit discrepancies in either quantity or sales value, with some instances showing up to 4 different values.

2.4 Fields that are Challenging to Understand

- We noticed that the same `RECEIPT_ID` and `BARCODE` sometimes show different `FINAL_QUANTITY`, `FINAL_SALE`, or both. The reasons behind these inconsistencies aren't entirely clear, but a few possibilities include:
    - **Scanning Errors** – Misreads during checkout could lead to incorrect quantity or pricing.
    - **System Processing Issues** – Data retrieval or parsing errors might cause duplicate or altered records.
    - **Retail Pricing Factors** – Discounts, bulk purchases, or promotions such as "Buy 1 Get 1 Free" could impact how quantities and sales values are recorded.

Additionally, it's unclear why some `FINAL_QUANTITY` values include zeros or decimal points, and why `FINAL_SALE` sometimes contains blank spaces. Further investigation is needed to determine if these are errors or intentional system behaviors.

2.4 Data Cleaning Approach Before Analysis

- Removed duplicates from `Products` datasets; treated duplicate entries as multiple scanned items in transaction datasets.
- Replaced `zero` values in `FINAL_QUANTITY` and blank spaces in `FINAL_SALE` with `0`.
- Filtered out transactions without `BARCODE`.
- Consolidated duplicate entries in `TRANSACTION` for better analysis.

3. SQL Queries and Insights

ThSQL queries were performed to extract insights from the cleaned dataset:

3.1 Closed-Ended Questions

1. Top 5 Brands by Receipts Scanned (Users 21+)

```
SELECT p.BRAND, COUNT(t.RECEIPT_ID) AS receipt_count
FROM TRANSACTION_TAKEHOME t
JOIN USER_TAKEHOME u ON t.USER_ID = u.ID
JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE
WHERE u.BIRTH_DATE <= DATE('2025-02-19', '-21 years')  -- Users 21 and
older
AND p.BRAND IS NOT NULL
GROUP BY p.BRAND
ORDER BY receipt_count DESC
LIMIT 5;
```

Brand	Receipt Count
NERDS CANDY	6
DOVE	6

TRIDENT	4
SOUR PATCH KIDS	4
MEIJER	4

2. Top 5 Brands by Sales (Users with 6+ Months of Account Tenure)

```
SELECT p.BRAND, SUM(CAST(t.true_FINAL_SALE AS FLOAT)) 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 u.CREATED_DATE <= DATE('2025-02-19', '-6 months')  -- Users with at
least six months tenure
AND p.BRAND IS NOT NULL
GROUP BY p.BRAND
ORDER BY total_sales DESC
LIMIT 5;
```

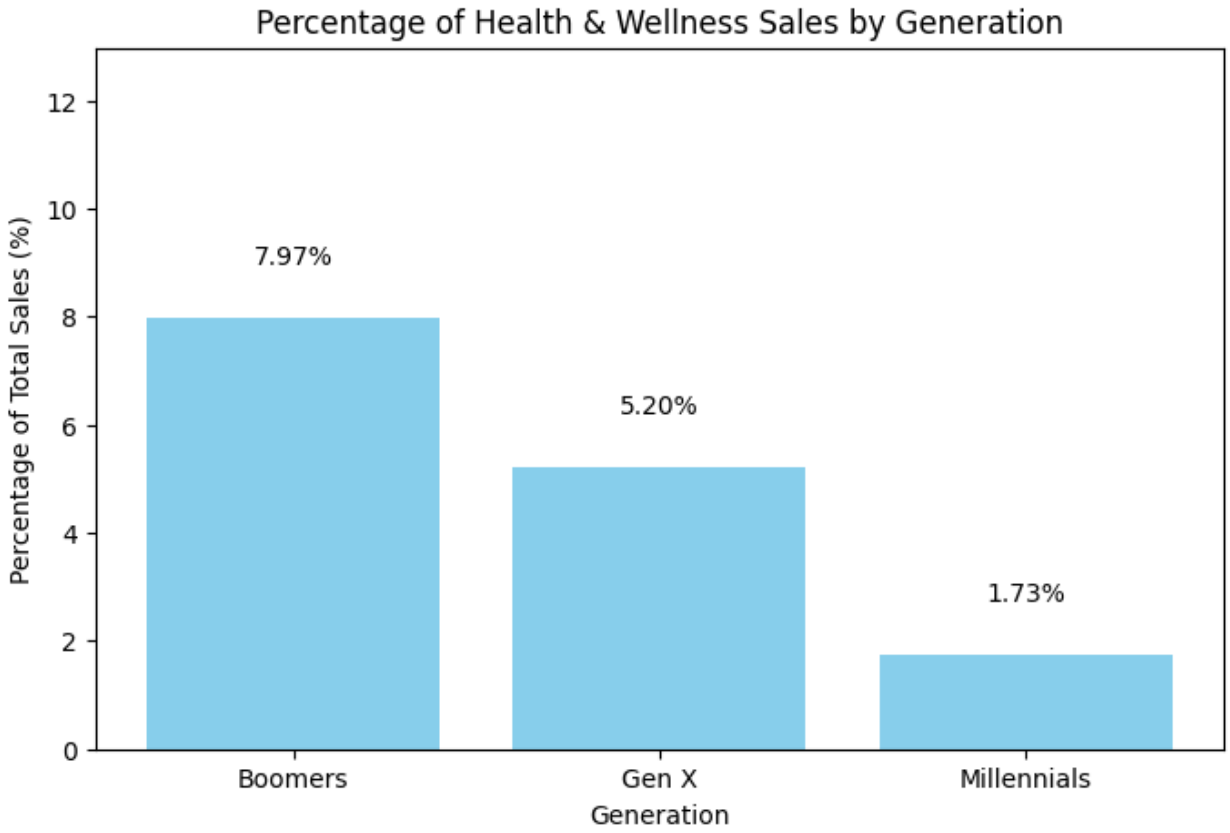
Brand	Total Sales ($)
CVS	72.00
DOVE	30.91
TRESEMMÉ	29.16
TRIDENT	23.36
COORS LIGHT	17.48

3. Percentage of Health & Wellness Sales by Generation

```
SELECT
    CASE
        WHEN DATE('2025-03-03', '-25 years') <= u.BIRTH_DATE THEN 'Gen Z'
        WHEN DATE('2025-03-03', '-40 years') <= u.BIRTH_DATE THEN
'Millennials'
        WHEN DATE('2025-03-03', '-55 years') <= u.BIRTH_DATE THEN 'Gen X'
        ELSE 'Boomers'
    END AS generation,
    SUM(t.true_FINAL_SALE) AS total_health_sales,
    (SUM(t.true_FINAL_SALE) * 100.0 ) / (SELECT SUM(true_FINAL_SALE) FROM
TRANSACTION_TAKEHOME) AS percentage_of_total
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;
```

Generation	Total Sales ($)	Percentage of Total Sales
Boomers	93.13	7.97%

Gen X	60.78	5.20%
Millennials	20.21	1.73%



3.2 Open-Ended Questions

1. Fetch's Power Users

Power users were defined as those with the highest number of receipts scanned. The top 10 users had between 12 and 20 receipts each. Interestingly, most of these users were absent in the `USER dataset`, indicating potential data linkage issues.

```
SELECT t.USER_ID, COUNT(t.RECEIPT_ID) AS receipt_count
FROM TRANSACTION_TAKEHOME t
GROUP BY t.USER_ID
ORDER BY receipt_count DESC
LIMIT 10;
```

	USER_ID	receipt_count
0	62925c1be942f00613f7365e	20
1	64063c8880552327897186a5	18
2	6327a07aca87b39d76e03864	14
3	61d5f5d2c4525a3a478b386b	14
4	60a5363facc00d347abadc8e	14
5	609af341659cf474018831fb	14
6	6682cbf6465f309038ae1888	12
7	66651af0e04f743a096e3bf9	12
8	653a0f40909604bae9071473	12
9	63f1904938f010745b9a2b60	12

2. Leading Brand in Dips & Salsa Category

A SQL query identified the top-selling brand in this category, though further analysis is required to validate trends across different periods.

```
SELECT p.BRAND, SUM(CAST(t.true_FINAL_SALE AS FLOAT)) AS total_sales
FROM TRANSACTION_TAKEHOME t
JOIN PRODUCTS_TAKEHOME p ON t.BARCODE = p.BARCODE
WHERE p.CATEGORY_2 = 'Dips & Salsa'
GROUP BY p.BRAND
ORDER BY total_sales DESC
LIMIT 1;
```

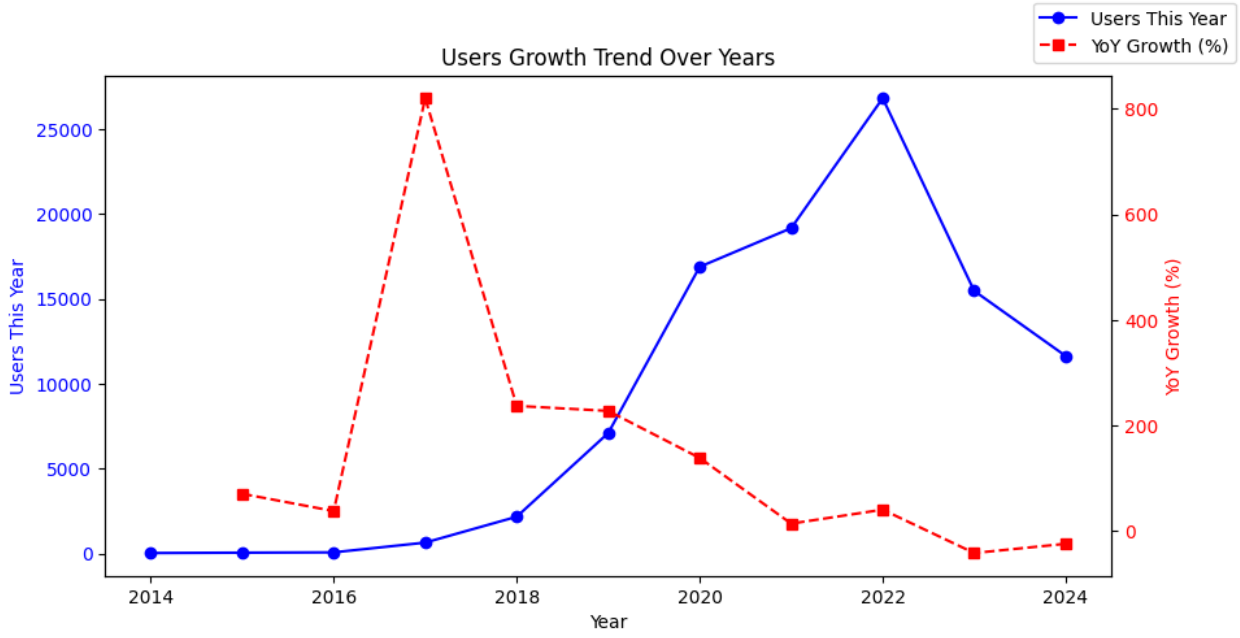
	BRAND	total_sales
0	TOSTITOS	197.24

3. Fetch's Year Growth Percentage

```
WITH UserYearly AS (
    SELECT
        CAST(strftime('%Y', CREATED_DATE) AS INTEGER) AS Year,
        COUNT(ID) AS NewUsers
    FROM USER_TAKEHOME
    WHERE CREATED_DATE IS NOT NULL
    GROUP BY Year
)
SELECT
```

```
    U1.Year AS Current_Year,
    U1.NewUsers AS Users_This_Year,
    U2.NewUsers AS Users_Last_Year,
    ROUND(((U1.NewUsers - U2.NewUsers) * 100.0 / NULLIF(U2.NewUsers, 0)),
2) AS YoY_Growth_Percentage
FROM UserYearly U1
LEFT JOIN UserYearly U2
    ON U1.Year = U2.Year + 1  -- Ensure correct year pairing
ORDER BY U1.Year DESC;
```

	Current_Year	Users_This_Year	Users_Last_Year	YoY_Growth_Percentage
0	2024	11631	15464.0	-24.79
1	2023	15464	26807.0	-42.31
2	2022	26807	19159.0	39.92
3	2021	19159	16883.0	13.48
4	2020	16883	7093.0	138.02
5	2019	7093	2168.0	227.17
6	2018	2168	644.0	236.65
7	2017	644	70.0	820.00
8	2016	70	51.0	37.25
9	2015	51	30.0	70.00
10	2014	30	NaN	NaN



4. Communication to Stakeholders

Hi Team,

We've completed an initial analysis of Fetch's transactional and user data, and we've identified key areas that need attention:

1. Data Quality Issues:

- High missing values in `LANGUAGE`, `CATEGORY_4`, `MANUFACTURER`, and `BRAND`, which impact product-level insights.
- Inconsistencies in `FINAL_QUANTITY` and `FINAL_SALE` fields, where multiple records for the same transaction show different values.
- Over 50% of transactions have discrepancies in recorded quantity or sales price, suggesting potential scanning or processing errors.

2. Key Trend Identified:

- **Walmart** has the highest total sales but also shows significant transaction inconsistencies, which might indicate reporting or pricing anomalies.
- **Boomers** account for the highest percentage (7.97%) of sales in the Health & Wellness category, a valuable insight for targeted marketing strategies.
- **Power users** are not logged in our user dataset; this might lead to potential customer loss in the long run.
- **Declining** user acquisition in 2023 (-42%) and 2024 (-24%) suggests reassessing marketing and engagement strategies to reverse the trend.

3. Recommended Actions:

- **Improve Data Integrity:** Standardize how `FINAL_QUANTITY` and `FINAL_SALE` values are recorded.
- **Address Data Gaps:** Investigate why high-value users (power users) are missing from the user dataset.
- **Refine Transaction Logging:** Verify whether promotional discounts and system errors are affecting recorded sales values.
- **Enhance Barcode Scanning Accuracy:** Minimize errors in duplicate or inconsistent barcode transactions.
- **Improve User Retention:** Analyze user engagement data to enhance retention and reduce churn.

Let's discuss these findings further and decide the next steps. Looking forward to your insights.

Best,

Angel

5. Conclusion

This analysis has identified key data quality issues, notable trends, and areas for improvement in Fetch's transactional and user datasets. Significant missing values, duplicate records, and inconsistencies in transaction data indicate a need for improved data integrity and processing standards. Analyzing Fetch's power users and generational spending patterns provides actionable insights for customer engagement and marketing strategies. Additionally, the decline in user acquisition in 2023 and 2024 highlights an urgent need to reassess growth strategies. Fetch should prioritize improving data completeness, refining transaction logging mechanisms, and implementing targeted user retention and acquisition strategies to enhance data reliability and business performance. Addressing these areas will ensure better decision-making, more substantial customer insights, and sustained business growth.