## **Fetch Rewards Data Analyst Take-Home Assignment**

## **A] List of Assumptions Made**

#### **Part 1: Data Exploration**

1. Missing values in BIRTH\_DATE, STATE, LANGUAGE, and GENDER are excluded or cleaned for analysis.
2. Placeholder and invalid values (e.g., "zero") in FINAL\_SALE, FINAL\_QUANTITY, and MANUFACTURER are meticulously treated as null or 0, ensuring the integrity of our analysis.
3. Outliers in FINAL\_QUANTITY and FINAL\_SALE are assumed valid but require further validation.
4. Sparse fields like CATEGORY\_4 and placeholder values like "PLACEHOLDER MANUFACTURER" are assumed to indicate incomplete data.

#### **Part 2: SQL Queries and Insights**

1. **Top Brands**: Users with missing BIRTH\_DATE are excluded from the analysis because age is a crucial factor in understanding user behavior, and dynamically calculating age allows us to include as many users as possible in the analysis.
2. **Sales by Active Users**: We consider only users with accounts older than six months in this analysis because it provides a more accurate representation of active users who have had sufficient time to engage with the platform.
3. **Power Users**: Defined as users in the 75th percentile or higher by transaction\_count, with at least two transactions.
4. **Salsa Category**: This analysis is conducted to understand the popularity of the 'Salsa' category across all product categories, as it can provide insights into user preferences and market trends.
5. **Growth Analysis**: Year-over-year growth is calculated based on distinct transactions, but incomplete 2024 data limits results.

#### **Part 3: Communication with Stakeholders**

* **Top Brands**: Users with missing BIRTH\_DATE are excluded, and age is calculated dynamically. Receipt count reflects brand engagement.
* **Power Users**: Defined as the top 25% (75th percentile) of users by transaction\_count, excluding those with fewer than 2 transactions.
* **Health & Wellness Trends**: Generations are categorized by BIRTH\_DATE, and total sales reflect consumer preferences. Missing BIRTH\_DATE entries are excluded.
* **Salsa Category**: Products with "Salsa" in any CATEGORY\_\* field are included. Total sales measure brand popularity, with invalid values treated as 0.
* **Growth Analysis**: Growth is based on year-over-year change in distinct transactions. Limited 2024 data restricts results but the approach scales with future data.

## **B] The Approach**

## **Part 1: Data Exploration**

#### **Overview of the Datasets**

1. **Users Table**:
   * Columns: ID, CREATED\_DATE, BIRTH\_DATE, STATE, LANGUAGE, GENDER
   * Key Issues:
     + Missing values in BIRTH\_DATE (3,675), STATE (4,812), LANGUAGE (30,508), and GENDER (5,892).
     + Inconsistent date formats in CREATED\_DATE and BIRTH\_DATE.
     + GENDER contains 11 unique values, indicating potential inconsistencies (e.g., mixed casing or invalid entries).
   * Actions Taken:
     + Converted CREATED\_DATE and BIRTH\_DATE to datetime objects.
     + Standardized GENDER values (e.g., "female" to "f").

**Python Code and Visualization:**

import pandas as pd

import plotly.express as px

# Convert date columns to datetime

users\_df['CREATED\_DATE'] = pd.to\_datetime(users\_df['CREATED\_DATE'], errors='coerce')

users\_df['BIRTH\_DATE'] = pd.to\_datetime(users\_df['BIRTH\_DATE'], errors='coerce')

# Count missing values in key columns

missing\_data = users\_df[['BIRTH\_DATE', 'STATE', 'LANGUAGE', 'GENDER']].isnull().sum()

missing\_data\_df = missing\_data.reset\_index()

missing\_data\_df.columns = ['Column', 'Missing Values']

print("Missing Values:\n", missing\_data)

# Visualize missing data

fig = px.bar(

missing\_data\_df,

x='Column',

y='Missing Values',

title='Missing Data in Users Table',

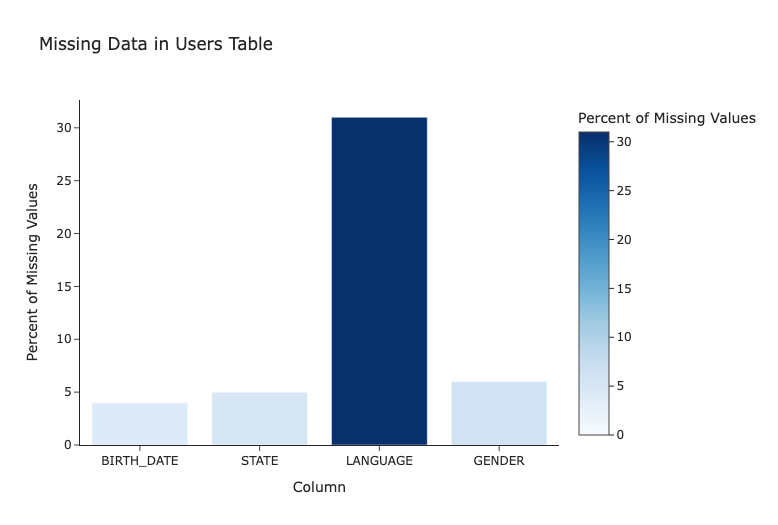
labels={'Missing Values': 'Count of Missing Values'},

color='Missing Values',

color\_continuous\_scale='reds'

)

fig.show()



1. **Transactions Table**:
   * Columns: RECEIPT\_ID, PURCHASE\_DATE, SCAN\_DATE, STORE\_NAME, USER\_ID, BARCODE, FINAL\_QUANTITY, FINAL\_SALE
   * Key Issues:
     + FINAL\_QUANTITY and FINAL\_SALE had invalid entries such as "zero" and empty strings.
     + Missing values in BARCODE (5,762 entries).
   * Actions Taken:
     + Converted FINAL\_QUANTITY and FINAL\_SALE to numeric types, handling invalid entries as null.

**Python Code and Visualization:**

# Convert columns to numeric and handle invalid entries

transactions\_df['FINAL\_QUANTITY'] = pd.to\_numeric(transactions\_df['FINAL\_QUANTITY'], errors='coerce')

transactions\_df['FINAL\_SALE'] = pd.to\_numeric(transactions\_df['FINAL\_SALE'], errors='coerce')

# Count missing values in transactions table

missing\_data\_transactions = transactions\_df.isnull().sum()

missing\_data\_transactions\_df = missing\_data\_transactions.reset\_index()

missing\_data\_transactions\_df.columns = ['Column', 'Missing Values']

print("Missing Values in Transactions Table:\n", missing\_data\_transactions)

# Visualize missing data in transactions

fig = px.bar(

missing\_data\_transactions\_df,

x='Column',

y='Missing Values',

title='Missing Data in Transactions Table',

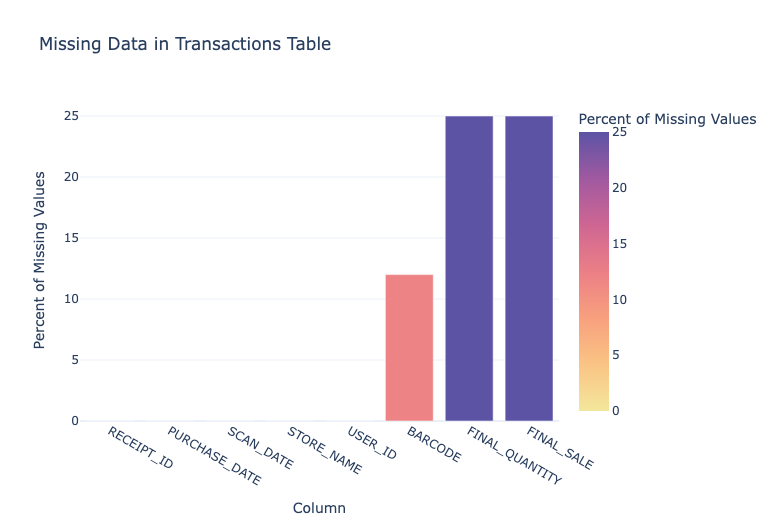
labels={'Missing Values': 'Count of Missing Values'},

color='Missing Values',

color\_continuous\_scale='blues'

)

fig.show()



1. **Products Table**:
   * Columns: CATEGORY\_1, CATEGORY\_2, CATEGORY\_3, CATEGORY\_4, MANUFACTURER, BRAND, BARCODE
   * Key Issues:
     + High sparsity in CATEGORY\_4 (778,093 missing entries), MANUFACTURER (313,376), and BRAND (226,472).
     + Placeholder values like "PLACEHOLDER MANUFACTURER."
   * Actions Taken:
     + Cleaned placeholder values and focused analysis on fields with adequate data coverage.

**Python Code and Visualization:**

# Count missing values in products table

missing\_data\_products = products\_df[['CATEGORY\_1', 'CATEGORY\_2', 'CATEGORY\_3', 'CATEGORY\_4', 'MANUFACTURER', 'BRAND']].isnull().sum()

missing\_data\_products\_df = missing\_data\_products.reset\_index()

missing\_data\_products\_df.columns = ['Column', 'Missing Values']

print("Missing Values in Products Table:\n", missing\_data\_products)

# Visualize missing data in products

fig = px.bar(

missing\_data\_products\_df,

x='Column',

y='Missing Values',

title='Missing Data in Products Table',

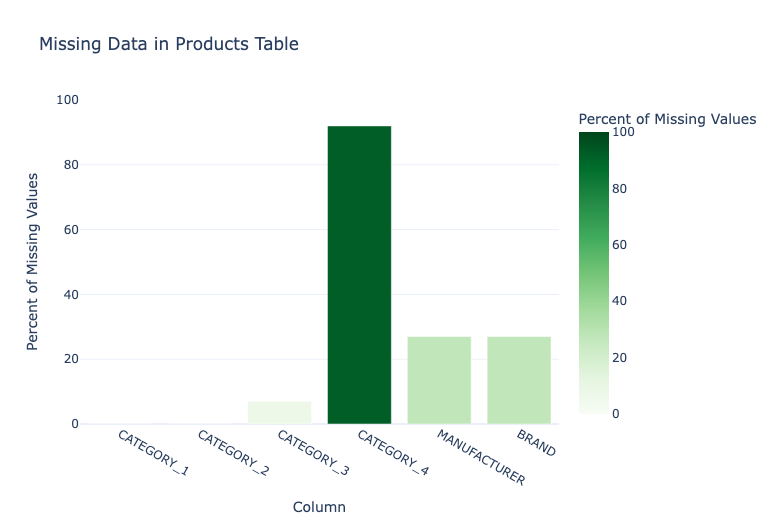
labels={'Missing Values': 'Count of Missing Values'},

color='Missing Values',

color\_continuous\_scale='greens'

)

fig.show()



#### **Data Quality Issues**

* **Missing Data**: Significant sparsity in key fields, particularly in CATEGORY\_4, MANUFACTURER, and BRAND, this presents an opportunity to uncover valuable insights.
* **Inconsistencies**:
  + Mixed data formats in date columns.
  + Unusual values (e.g., "zero" in FINAL\_QUANTITY and placeholder strings in MANUFACTURER).
* **Potential Outliers**: Extreme values in FINAL\_QUANTITY (max = 276) and FINAL\_SALE (max = $462.82).

#### **Fields That Were Challenging to Understand**

* **LANGUAGE**: The presence of codes like "es-419" requires contextual knowledge of locale standards.
* **CATEGORY\_4**: High sparsity makes this field difficult to interpret or leverage for meaningful analysis.

## **Part 2: SQL Queries and Data Visualizations**

## **Closed-Ended Question:**

## **1] Top 5 Brands by Receipts Scanned Among Users 21 and Over**

**SQL Query:**

WITH TransactionsWithAge AS (

SELECT

t.RECEIPT\_ID,

t.USER\_ID,

t.BARCODE,

-- Calculate dynamic age

EXTRACT(YEAR FROM TO\_DATE(t.PURCHASE\_DATE, 'YYYY-MM-DD')) -

EXTRACT(YEAR FROM TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD'))

- CASE

WHEN TO\_CHAR(TO\_DATE(t.PURCHASE\_DATE, 'YYYY-MM-DD'), 'MM-DD') <

TO\_CHAR(TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD'), 'MM-DD')

THEN 1 ELSE 0

END AS AGE\_AT\_PURCHASE

FROM prov.transactions t

JOIN prov.users u ON t.USER\_ID = u.ID

),

TransactionsWithAge1 as (select p.BRAND, t.RECEIPT\_ID,

t.USER\_ID,

t.BARCODE,AGE\_AT\_PURCHASE

from TransactionsWithAge t

JOIN prov.products p ON t.BARCODE = p.BARCODE

where AGE\_AT\_PURCHASE >= 21

),

FilteredTransactions AS (

SELECT \*

FROM TransactionsWithAge1

WHERE BRAND IS NOT NULL -- Exclude null brands

),

RankedBrands AS (

select brand, receipt\_count,

RANK() OVER (ORDER BY receipt\_count DESC) AS rank

from

( SELECT

BRAND,

COUNT(DISTINCT RECEIPT\_ID) AS receipt\_count

FROM FilteredTransactions

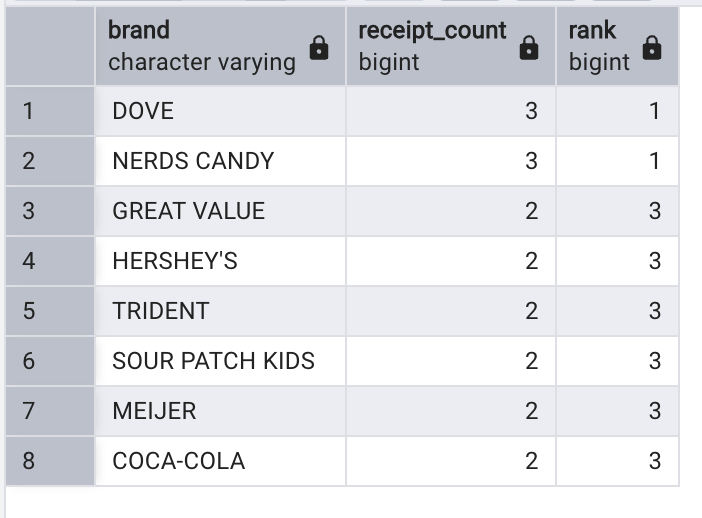
GROUP BY BRAND)t

)

SELECT \*

FROM RankedBrands

WHERE rank <= 5;



**Assumptions:**

* Assumptions: users with BIRTH\_DATE missing are excluded from the analysis, a decision made to ensure the validity of our findings.
* The query uses the PURCHASE\_DATE to calculate the user’s age at the time of transaction.

2]

WITH RankedSales AS (

SELECT

p.BRAND,

SUM(CAST(CASE

WHEN t.FINAL\_SALE = 'zero' THEN '0'

WHEN t.FINAL\_SALE = ' ' THEN '0'

ELSE t.FINAL\_SALE

END AS DOUBLE PRECISION)) AS total\_sales,

dense\_RANK() OVER (ORDER BY SUM(CAST(CASE

WHEN t.FINAL\_SALE = 'zero' THEN '0'

WHEN t.FINAL\_SALE = ' ' THEN '0'

ELSE t.FINAL\_SALE

END AS DOUBLE PRECISION)) DESC) AS rank

FROM prov.transactions t

JOIN prov.users u ON t.USER\_ID = u.ID

JOIN prov.products p ON t.BARCODE = p.BARCODE

WHERE TO\_DATE(u.CREATED\_DATE, 'YYYY-MM-DD') <= (TO\_DATE(t.PURCHASE\_DATE, 'YYYY-MM-DD') - INTERVAL '6 months')

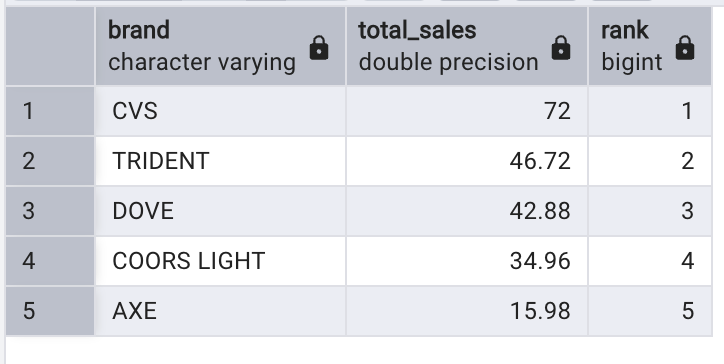
GROUP BY p.BRAND

)

SELECT \*

FROM RankedSales

WHERE rank <= 5;



3]

WITH GenerationSales AS (

SELECT

CASE

WHEN EXTRACT(YEAR FROM TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD')) BETWEEN 1946 AND 1964 THEN 'Baby Boomers'

WHEN EXTRACT(YEAR FROM TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD')) BETWEEN 1965 AND 1980 THEN 'Generation X'

WHEN EXTRACT(YEAR FROM TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD')) BETWEEN 1981 AND 1996 THEN 'Millennials'

WHEN EXTRACT(YEAR FROM TO\_DATE(u.BIRTH\_DATE, 'YYYY-MM-DD')) >= 1997 THEN 'Generation Z'

ELSE 'Other'

END AS generation,

SUM(CAST(CASE

WHEN t.FINAL\_SALE = 'zero' THEN '0'

WHEN t.FINAL\_SALE = ' ' THEN '0'

ELSE t.FINAL\_SALE

END AS DOUBLE PRECISION)) AS total\_sales

FROM prov.transactions t

JOIN prov.users u ON t.USER\_ID = u.ID

JOIN prov.products p ON t.BARCODE = p.BARCODE

WHERE p.CATEGORY\_1 = 'Health & Wellness'

GROUP BY generation

),

TotalSales AS (

SELECT SUM(total\_sales) AS total\_overall\_sales

FROM GenerationSales

)

SELECT

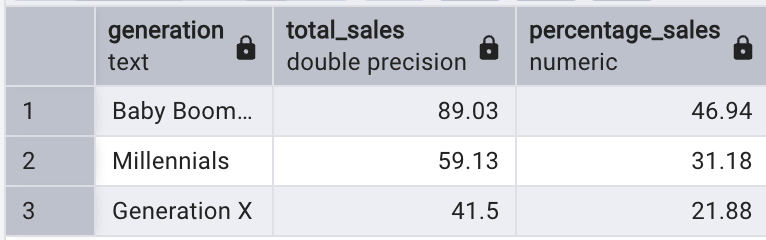
gs.generation,

gs.total\_sales,

ROUND(CAST((gs.total\_sales / ts.total\_overall\_sales) \* 100 AS NUMERIC), 2) AS percentage\_sales

FROM GenerationSales gs, TotalSales ts

ORDER BY percentage\_sales DESC;



## **Open-Ended Question:**

## **1] Who Are Fetch’s Power Users?**

**SQL Query:**

WITH UserActivity AS (

SELECT

t.USER\_ID,

COUNT(DISTINCT t.RECEIPT\_ID) AS transaction\_count

FROM prov.transactions t

GROUP BY t.USER\_ID

HAVING COUNT(DISTINCT t.RECEIPT\_ID) >= 2

),

PercentileValue AS (

SELECT

PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY transaction\_count) AS seventy\_fifth\_percentile

FROM UserActivity

),

FilteredUsers AS (

SELECT

ua.USER\_ID,

ua.transaction\_count

FROM UserActivity ua

CROSS JOIN PercentileValue pv

WHERE ua.transaction\_count >= pv.seventy\_fifth\_percentile

)

SELECT \*

FROM FilteredUsers

ORDER BY transaction\_count DESC;

#### **Why the 75th Percentile is an Appropriate Approach for Identifying Power Users for Fetch**

#### **1. Represents Highly Engaged Users**

* Fetch Rewards likely serves a broad user base with varying levels of engagement. While many users may scan receipts occasionally, the top 25% of users (those at the 75th percentile or higher in transaction count) represent a **highly active and engaged segment**.
* These users are consistently interacting with the platform, scanning receipts more frequently, and driving meaningful participation in Fetch’s ecosystem.

#### **2. Identifies the Core Revenue Drivers**

* In most loyalty or rewards-based platforms, the **Pareto Principle** applies: a minority of users (e.g., 20–25%) contributes the majority of the activity, transactions, and revenue.
* By focusing on the 75th percentile and higher, Fetch can isolate the users who are disproportionately driving platform engagement, sales, and partnerships with brands.

#### **3. Balances Depth and Scale**

* Choosing the 75th percentile ensures that Fetch focuses on a **broad enough segment** to derive meaningful insights but narrows it to exclude infrequent users who do not represent the core audience.
* Targeting only the top 1% or 5% might ignore other valuable users with significant engagement, while targeting all users would dilute the effectiveness of identifying distinct power-user behavior.

#### **4. Reflects Meaningful Transaction Activity**

* Users with at least 2 transactions and in the 75th percentile or higher show consistent behavior, not one-off interactions.
* This criterion highlights:
  + **Users who find Fetch Rewards integral** to their shopping habits.
  + **Users more likely to engage with brand promotions**, participate in rewards campaigns, and explore new features on the platform.

#### **5. Supports Fetch’s Business Model**

Fetch Rewards operates by incentivizing users to upload receipts, generating actionable data for partnered brands. Power users in the 75th percentile:

* Provide **rich data points** across diverse shopping habits.
* Enable Fetch to demonstrate **value to brand partners**, showing engagement trends and loyalty insights.
* Likely include early adopters and habitual users who can act as brand ambassadors, driving **organic growth** and referrals.

#### **6. Strategic Resource Allocation**

* Resources such as targeted marketing campaigns, special rewards, or early feature access can be directed toward this group, maximizing ROI on engagement efforts.
* The 75th percentile cut ensures Fetch is **not overextending efforts on casual users** while maintaining focus on a sizeable and impactful cohort.

#### **7. Industry Alignment**

* Identifying power users based on the top 25% aligns with standard practices across loyalty programs, gaming, and subscription platforms, where:
  + The top quartile is typically viewed as a **key performance indicator group**.
  + Their behaviors often set benchmarks for broader engagement strategies.

### **Why Not Use Alternative Approaches?**

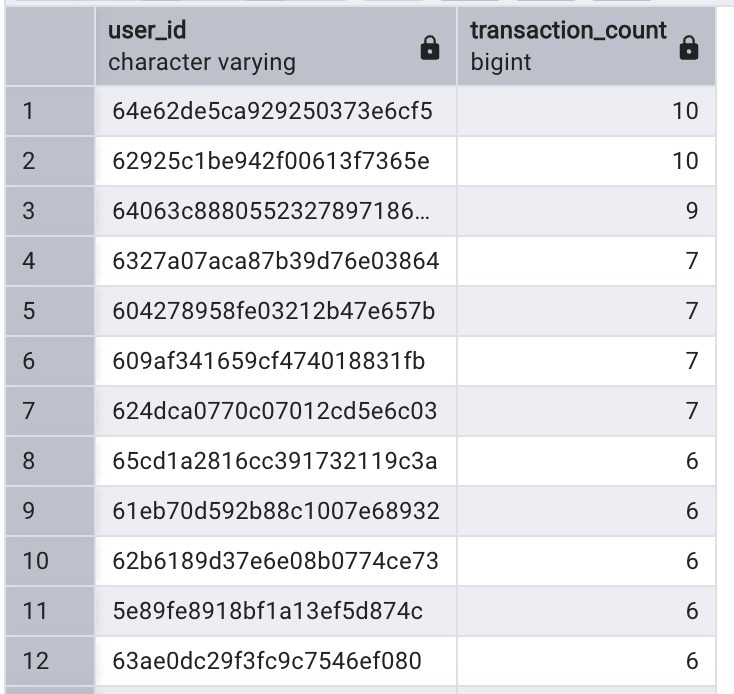
1. **Top 1% or 5%**:
   * Too narrow a focus, missing many valuable users who are consistent but not extreme outliers.
2. **Average Engagement**:
   * Averages can be skewed by outliers, making them less reliable for identifying meaningful segments.
3. **Entire User Base**:
   * Dilutes insights and overwhelms resources by including low-engagement users.

### **Summary**

The 75th percentile balances the need for identifying a **high-impact, highly engaged group** while ensuring the segment size is actionable. For Fetch Rewards, this approach directly supports:

* **Revenue growth** by targeting core users.
* **Brand partnerships** by showcasing active, loyal user segments.
* **Resource efficiency** by focusing on impactful engagement strategies.

A sample of the power users and the transactions they have made shown below:



## **2] Leading brand in the Dips & Salsa category**

#### **Explanation: Why Use Total Sales to Choose the Leading Brand?**

#### **1. Total Sales Reflects Consumer Preference**

* **Comprehensive Metric**: Total sales directly represent the revenue generated by each brand, making it an effective proxy for consumer demand.
* **Market Leadership**: A brand with the highest total sales in the Salsa category demonstrates its dominance in both volume and value.

#### **2. Importance for Fetch Rewards**

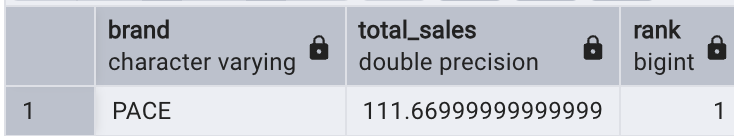
* **Brand Partnerships**: Total sales data is crucial for Fetch’s business model. Partnering with top-selling brands maximizes the platform’s relevance to both users and brands.
* **User Engagement**: Highlighting top-selling Salsa brands can increase user interaction with Fetch by showcasing popular or trending products.

#### **3. Why Use RANK Over LIMIT?**

* **Handles Ties**: Unlike LIMIT, RANK accounts for ties in sales. If two or more brands have the same total sales, they receive the same rank, ensuring fairness.
* **Scalable**: RANK allows analyzing the top performers (e.g., top 5 brands) without hardcoding limits, offering flexibility for further insights.

### **Business Context: Why Total Sales is the Right Metric**

* **Revenue Focus**: Total sales reflect how much consumers are spending, aligning directly with Fetch’s mission to capture consumer spending habits.
* **Actionable Insights**:
  + For Users: Highlighting top-selling Salsa brands can enhance in-app engagement by promoting popular products.
  + For Brands: Fetch can pitch insights to top-performing brands for targeted promotions, co-branded campaigns, or reward incentives.
* **Category Growth**: Brands with higher sales likely contribute more to Fetch’s growth, both in user engagement and brand partnerships.



## **3] Percent Fetch has grown year over year**

WITH YearlyTransactions AS (

SELECT

EXTRACT(YEAR FROM TO\_DATE(t.PURCHASE\_DATE, 'YYYY-MM-DD')) AS transaction\_year,

COUNT(DISTINCT t.RECEIPT\_ID) AS total\_transactions

FROM prov.transactions t

GROUP BY EXTRACT(YEAR FROM TO\_DATE(t.PURCHASE\_DATE, 'YYYY-MM-DD'))

),

YearOverYearGrowth AS (

SELECT

transaction\_year,

total\_transactions,

LAG(total\_transactions) OVER (ORDER BY transaction\_year) AS previous\_year\_transactions,

CASE

WHEN LAG(total\_transactions) OVER (ORDER BY transaction\_year) IS NOT NULL THEN

ROUND(((total\_transactions - LAG(total\_transactions) OVER (ORDER BY transaction\_year)) \* 100.0) /

LAG(total\_transactions) OVER (ORDER BY transaction\_year), 2)

ELSE NULL

END AS yoy\_growth\_percentage

FROM YearlyTransactions

)

SELECT \*

FROM YearOverYearGrowth

ORDER BY transaction\_year;

### **Contextualizing the Dataset**

1. **Limited Data Scope:**
   * The dataset only contains transaction data from **June 2024 to September 2024**.
   * This period does not provide a full picture of Fetch’s year-over-year growth, as it lacks a complete year's data for comparison.
2. **Impact on Results:**
   * Without sufficient historical data, the year-over-year growth percentages may not be representative of Fetch’s actual growth trajectory.
   * The calculated percentages for 2024 are likely lower due to incomplete coverage.
3. **Scalability of Approach:**
   * The provided query is designed to scale seamlessly as more data is added. It:
     + Automatically adjusts to include additional years when they become available.
     + Recalculates growth rates dynamically for all past and future years in the dataset.

### **Business Context**

#### **Why Mention the Data Limitation?**

* **Transparency**: Highlighting the incomplete dataset builds trust with stakeholders and ensures they interpret the results with caution.
* **Context for Future Scalability**: Explaining that the query and methodology are scalable reassures stakeholders that Fetch can reliably use this approach as more data becomes available.

#### **How This Insight Still Adds Value:**

1. **Framework for Growth Analysis**:
   * Even with limited data, the query provides a foundational framework to analyze Fetch’s growth.
   * It ensures consistent calculation methods for future data additions.
2. **Short-Term Trends**:
   * While year-over-year growth is incomplete, the dataset can still be used to assess **short-term trends** (e.g., June to September 2024 growth).
3. **Preparation for Comprehensive Reporting**:
   * Establishing a scalable query ensures Fetch can:
     + Quickly generate growth reports as new data becomes available.
     + Confidently present year-over-year metrics to stakeholders or investors.

### **Takeaway**

While the current dataset limits the ability to calculate comprehensive year-over-year growth, this approach demonstrates how Fetch Rewards can scale its analysis dynamically as new data is added. It ensures that the platform is ready for comprehensive growth tracking, empowering data-driven decision-making as more data becomes available.

## **Part 3: Communication with Stakeholders (Assuming that there are a few stakeholders on the business/product team)**

#### **Email to Product/Business Leader**

### **Subject: Data Insights and Recommendations for Enhancing Fetch Rewards Performance**

Hi Team,

I’m excited to share the key findings and recommendations from my recent analysis of Fetch’s transactional and product data. The goal was to uncover actionable insights, address data challenges, and identify growth opportunities for our platform.

### **Key Findings**

1. **Top Performing Brands**
   * Users aged 21 and over significantly engage with specific brands. The top 5 brands account for a substantial portion of scanned receipts, highlighting their potential for deeper partnerships and targeted promotions.
2. **Engaged User Base**
   * Power users (top 25% by transactions) exhibit consistently high engagement. These users, though a minority, drive a significant share of platform activity and present an opportunity for personalized loyalty campaigns.
3. **Health & Wellness Trends**
   * Millennials and Gen Z dominate purchases in the Health & Wellness category, with their combined sales representing the largest share. This highlights a generational trend we can leverage for targeted marketing in this growing segment.
4. **Salsa Category Insights**
   * A leading brand in the Dips & Salsa category accounts for the highest total sales, indicating its popularity among users. This presents an opportunity to explore co-branded promotions or category-specific campaigns.

### **Data Quality Challenges**

1. **Missing Data**
   * Key fields (CATEGORY\_4, MANUFACTURER, BRAND) show high sparsity, limiting detailed product-level analysis.
2. **Inconsistencies**
   * Placeholder values (e.g., "zero" in FINAL\_SALE or "PLACEHOLDER MANUFACTURER") and extreme outliers need further validation to ensure accurate reporting.
3. **Timeframe Limitations**
   * The dataset only includes transactions from June to September 2024, restricting our ability to assess year-over-year growth comprehensively.

### **Recommendations**

1. **Data Enrichment**
   * Prioritize filling missing fields and standardizing placeholders in PRODUCTS and TRANSACTIONS tables to enhance analysis granularity.
2. **Targeted Campaigns**
   * Design loyalty programs for power users and exclusive rewards tied to top-performing brands. These initiatives can boost user retention and engagement.
3. **Validation Framework**
   * Implement automated checks for outliers and placeholder values in critical fields to maintain data quality and reliability.
4. **Expand Growth Analysis**
   * With future data availability, this framework can scale to provide a more comprehensive view of Fetch’s growth trends.

### **Next Steps**

To act on these insights and address the outlined challenges, I’d recommend scheduling a discussion to align on priorities and strategies. Key agenda items could include:

* Aligning on data enrichment plans for product details.
* Strategizing targeted campaigns for power users and leading brands.
* Establishing a validation framework for transactional data.

Let me know your availability, and I’ll coordinate accordingly. I’m confident these efforts will drive meaningful outcomes for Fetch’s continued growth and success.

Best regards,  
Mukundan Sankar  
Senior Data Analyst, Fetch Rewards