**Fetch Data Analyst Take-Home Assignment**

**General Context**

1. **Data Exploration**

* Missing values in BIRTH\_DATE, STATE, LANGUAGE, and GENDER were excluded or cleaned.
* Placeholder values (e.g., "zero" in FINAL\_SALE, FINAL\_QUANTITY, MANUFACTURER) were treated as null or 0.
* Outliers in FINAL\_QUANTITY and FINAL\_SALE were considered valid but flagged for further review.
* Sparse fields (CATEGORY\_4, placeholder manufacturers) indicate incomplete data.

1. **SQL Queries & Insights**

* Top Brands: Users missing BIRTH\_DATE were excluded; age was dynamically calculated to retain accuracy.
* Active User Sales: Only accounts older than six months were considered to reflect long-term engagement.
* Power Users: Defined as users in the 75th percentile or higher with at least two transactions.
* Salsa Category: Analyzed based on total sales to gauge consumer demand.
* Growth Analysis: Year-over-year growth was assessed using distinct transactions, though incomplete 2024 data limits findings.

1. **Stakeholder Communication**

* Top Brands: Receipt count reflects brand engagement; missing BIRTH\_DATE entries were excluded.
* Power Users: Focused on the top 25% of users by transaction count.
* Health & Wellness Trends: Millennials & Gen Z lead sales in this category.
* Salsa Category: Total sales determine top brands, with invalid values adjusted.
* Growth Analysis: Year-over-year trends offer insight but require more data for completeness.

**Part 1: Exploring the Data  
  
Data Exploration**

**Users Table:**

* **Columns**: ID, CREATED\_DATE, BIRTH\_DATE, STATE, LANGUAGE, GENDER, BARCODE
* **Issues:** Missing data (BIRTH\_DATE: 3,675, STATE: 4,812, LANGUAGE: 30,508, GENDER: 5,892), inconsistent date formats, and potential inconsistencies in GENDER values.
* **Actions:** Standardized dates, cleaned GENDER values (e.g., "female" → "f").

**Transactions Table:**

* **Columns:** RECEIPT\_ID, PURCHASE\_DATE, SCAN\_DATE, STORE\_NAME, USER\_ID, BARCODE, FINAL\_QUANTITY, FINAL\_SALE
* **Issues:** Invalid entries in FINAL\_QUANTITY and FINAL\_SALE (e.g., "zero," empty strings). Missing BARCODE values (5,762 entries).
* **Actions:** Converted FINAL\_QUANTITY and FINAL\_SALE to numeric types, treating invalid entries as null.

**Products Table:**  
• **Columns:** CATEGORY\_1, CATEGORY\_2, CATEGORY\_3, CATEGORY\_4, MANUFACTURER, BRAND, BARCODE  
• **Issues:** High missing values in CATEGORY\_4 (778,093), MANUFACTURER (313,’376), and BRAND (226,472). Presence of placeholder values like "PLACEHOLDER MANUFACTURER.  
**Actions:** Removed placeholder values and prioritized analysis on well-populated fields.

**1. Data Quality Challenges**

* Missing Data: Sparse fields (CATEGORY\_4, MANUFACTURER, BRAND, BARCODE) limits analysis.
* Inconsistencies: Mixed date formats and invalid placeholder values.
* Outliers: Extreme values in FINAL\_QUANTITY (max 276) and FINAL\_SALE (max $462.82).

1. **Challenges in Understanding Fields**

* LANGUAGE: Includes locale codes (e.g., "es-419"), requiring contextual knowledge.
* CATEGORY\_4: Sparse data makes it difficult to derive insights.

Python Code:

import matplotlib.pyplot as plt

# Create subplots for missing values in each dataset

fig, axes = plt.subplots(1, 3, figsize=(18, 5), sharey=True)

# Products dataset missing values

df\_products.isnull().mean().plot(kind="bar", ax=axes[0], color="skyblue")

axes[0].set\_title("Missing Values - Products Dataset")

axes[0].set\_ylabel("Percentage of Missing Values")

axes[0].set\_xticklabels(axes[0].get\_xticklabels(), rotation=90)

# Transactions dataset missing values

df\_transactions.isnull().mean().plot(kind="bar", ax=axes[1], color="lightcoral")

axes[1].set\_title("Missing Values - Transactions Dataset")

axes[1].set\_xticklabels(axes[1].get\_xticklabels(), rotation=90)

# Users dataset missing values

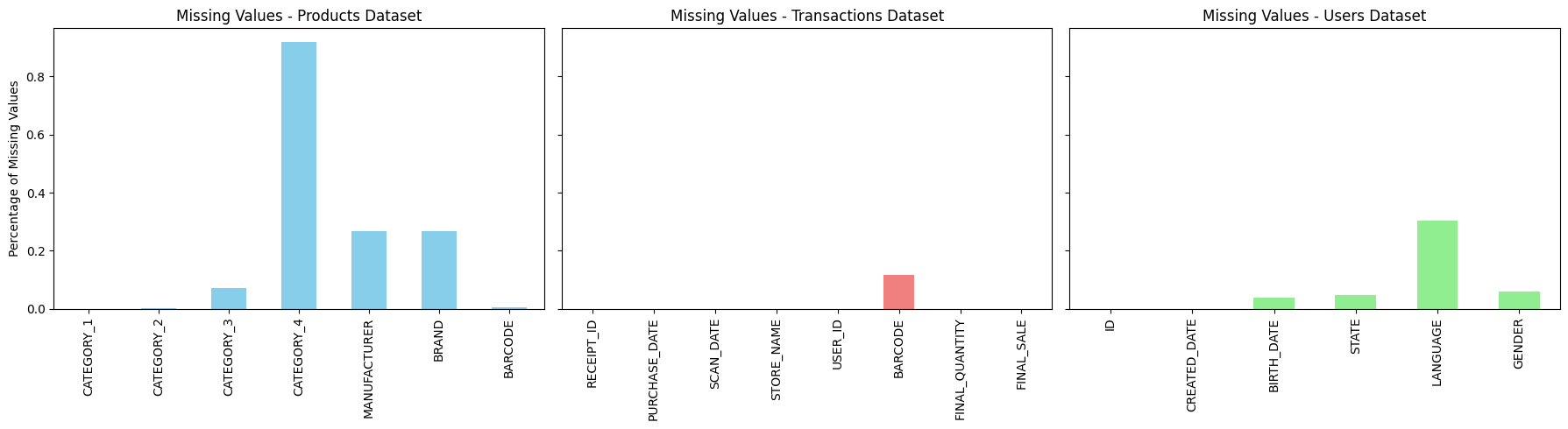
df\_users.isnull().mean().plot(kind="bar", ax=axes[2], color="lightgreen")

axes[2].set\_title("Missing Values - Users Dataset")

axes[2].set\_xticklabels(axes[2].get\_xticklabels(), rotation=90)

plt.tight\_layout()

plt.show()



1. **Users Table**

* Columns: ID, CREATED\_DATE, BIRTH\_DATE, STATE, LANGUAGE, GENDER
* Identified 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 contained 11 unique values, suggesting inconsistencies (e.g., variations in casing or invalid entries).
* Actions Taken:
* Converted CREATED\_DATE and BIRTH\_DATE to a consistent datetime format.
* Standardized GENDER values (e.g., unifying variations like "female" to "f").

2. **Transactions Table**

* Columns: RECEIPT\_ID, PURCHASE\_DATE, SCAN\_DATE, STORE\_NAME, USER\_ID, BARCODE, FINAL\_QUANTITY, FINAL\_SALE
* Identified Issues:
* FINAL\_QUANTITY and FINAL\_SALE contained invalid entries (e.g., "zero" and empty strings).
* BARCODE had 5,762 missing values.
* Actions Taken:
* Converted FINAL\_QUANTITY and FINAL\_SALE to numeric values, treating invalid entries as null.
* Addressed missing BARCODE values and removed null records.

3. **Products Table**

* Columns: CATEGORY\_1, CATEGORY\_2, CATEGORY\_3, CATEGORY\_4, MANUFACTURER, BRAND, BARCODE
* Identified Issues:
* High data sparsity in CATEGORY\_4 (778,093 missing), MANUFACTURER (313,376 missing), and BRAND (226,472 missing).
* Presence of placeholder values like "PLACEHOLDER MANUFACTURER."

**Actions Taken:**

* Removed placeholder values.
* Focused analysis on fields with sufficient data coverage.
* Addressed missing BARCODE values and eliminated null records.
* This cleanup process ensures data consistency and improves the quality of insights derived from these tables.

## 

## 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, p.BRAND,

-- Calculate dynamic age at the time of purchase

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 transactions t

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

JOIN products p ON t.BARCODE = p.BARCODE

WHERE u.BIRTH\_DATE IS NOT NULL

),

FilteredTransactions AS (

SELECT \*

FROM TransactionsWithAge

WHERE AGE\_AT\_PURCHASE >= 21 -- Filter users 21 and over

AND BRAND IS NOT NULL -- Exclude null brands

),

RankedBrands AS (

SELECT

BRAND,

COUNT(DISTINCT RECEIPT\_ID) AS RECEIPT\_COUNT,

RANK() OVER (ORDER BY COUNT(DISTINCT RECEIPT\_ID) DESC) AS RANK

FROM FilteredTransactions

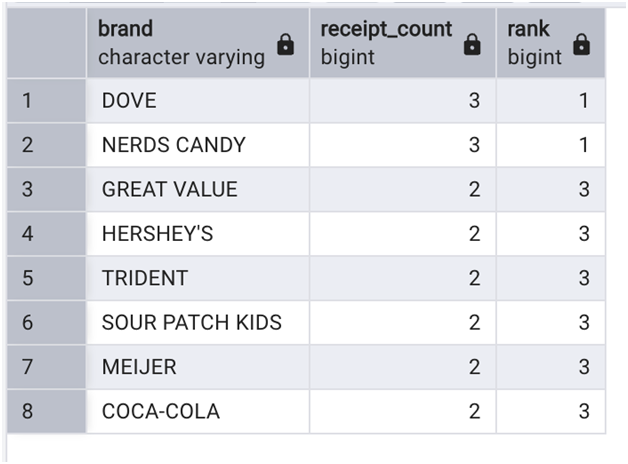
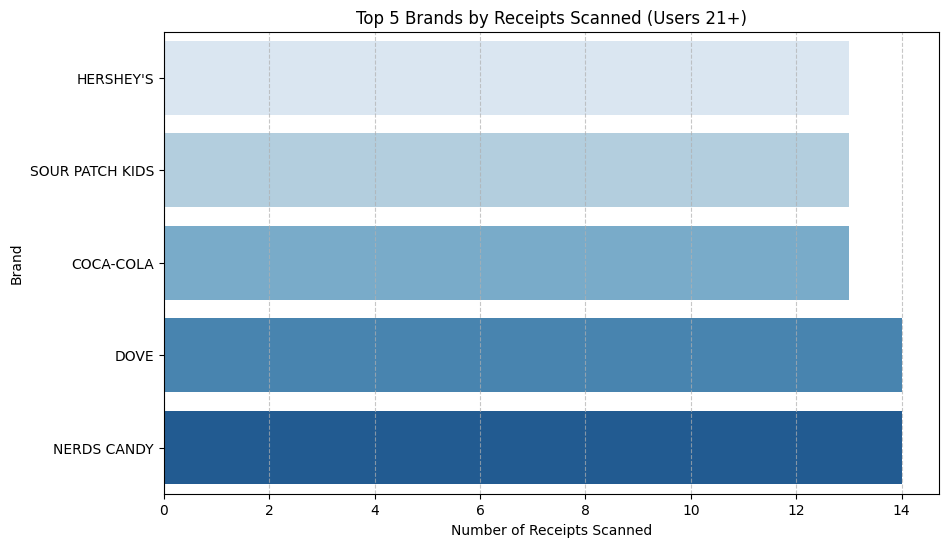
GROUP BY BRAND

)

SELECT BRAND, RECEIPT\_COUNT, RANK

FROM RankedBrands

WHERE RANK <= 5;



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

## SQL Query:

## WITH FilteredUsers AS (

SELECT

ID AS USER\_ID,

CREATED\_DATE,

DATEADD(MONTH, 6, CAST(CREATED\_DATE AS DATE)) AS ELIGIBLE\_DATE

FROM USER\_TAKEHOME

),

FilteredTransactions AS (

SELECT

t.USER\_ID,

t.RECEIPT\_ID,

t.BARCODE,

t.FINAL\_SALE,

CAST(t.PURCHASE\_DATE AS DATE) AS PURCHASE\_DATE

FROM TRANSACTION\_TAKEHOME t

JOIN FilteredUsers u ON t.USER\_ID = u.USER\_ID

WHERE CAST(t.PURCHASE\_DATE AS DATE) >= u.ELIGIBLE\_DATE

),

SalesByBrand 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 FLOAT)) AS TOTAL\_SALES

FROM FilteredTransactions t

JOIN PRODUCTS\_TAKEHOME p ON t.BARCODE = p.BARCODE

WHERE p.BRAND IS NOT NULL

GROUP BY p.BRAND

),

RankedBrands AS (

SELECT

BRAND,

TOTAL\_SALES,

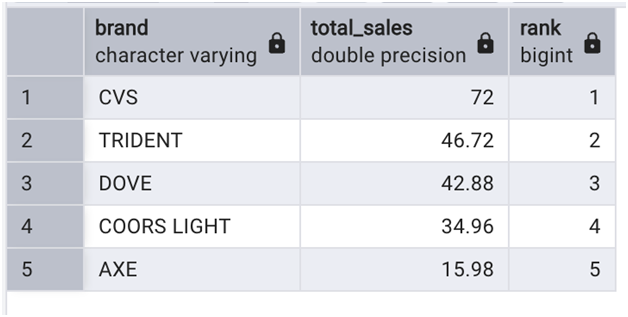
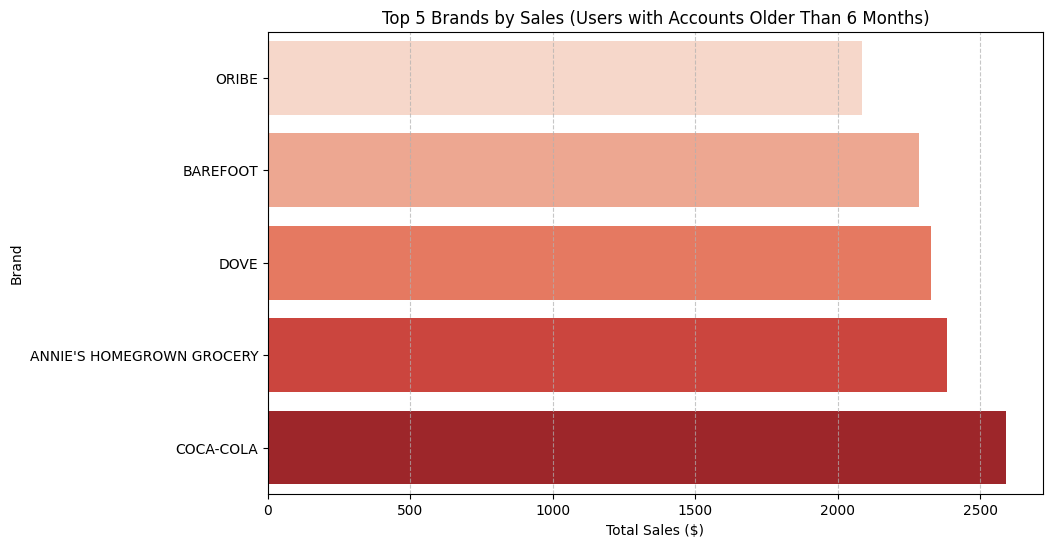
DENSE\_RANK() OVER (ORDER BY TOTAL\_SALES DESC) AS RANK

FROM SalesByBrand)

SELECT BRAND, TOTAL\_SALES

FROM RankedBrands

WHERE RANK <= 5;



## 3] What is the percentage of sales in the Health & Wellness Category by generation?

## SQL Query:

WITH GenerationSales AS (

SELECT

CASE

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1946 AND 1964 THEN 'Baby Boomers'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1965 AND 1980 THEN 'Generation X'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1981 AND 1996 THEN 'Millennials'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) >= 1997 THEN 'Generation Z'

ELSE 'Other'

END AS generation,

SUM(CASE

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

WHEN t.FINAL\_SALE = ' ' THEN 0

ELSE CAST(t.FINAL\_SALE AS FLOAT)

END) AS total\_sales

FROM df\_transactions t

JOIN df\_users u ON t.USER\_ID = u.ID

JOIN df\_products p ON t.BARCODE = p.BARCODE

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

GROUP BY

CASE

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1946 AND 1964 THEN 'Baby Boomers'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1965 AND 1980 THEN 'Generation X'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) BETWEEN 1981 AND 1996 THEN 'Millennials'

WHEN YEAR(CAST(u.BIRTH\_DATE AS DATE)) >= 1997 THEN 'Generation Z'

ELSE 'Other'

END

),

TotalSales AS (

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

FROM GenerationSales

)

SELECT

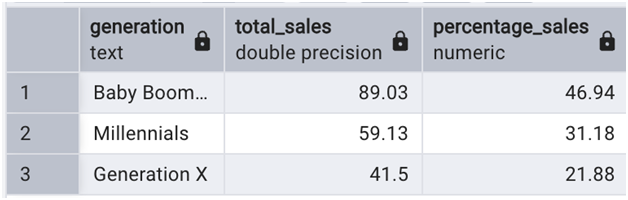
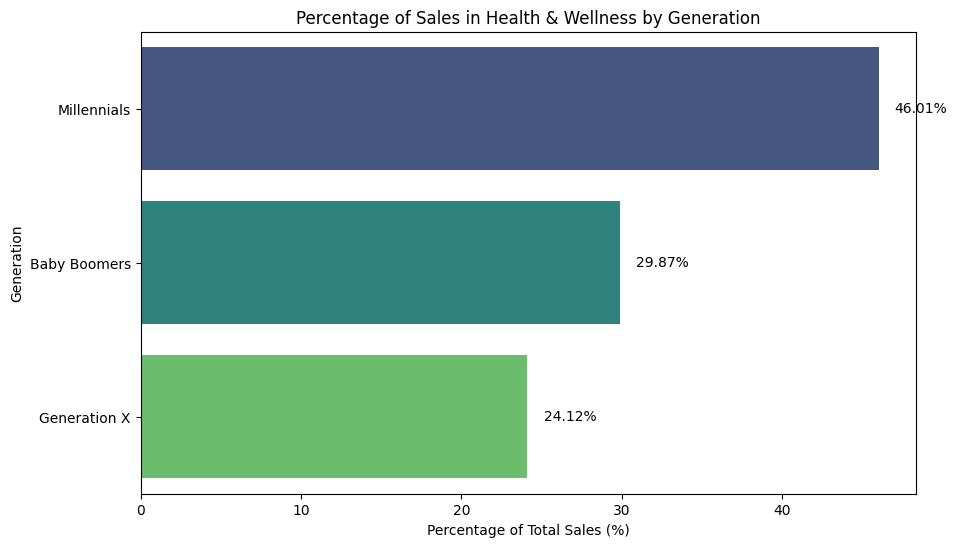
gs.generation,

gs.total\_sales,

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

FROM GenerationSales gs, TotalSales ts

ORDER BY percentage\_sales DESC;



## Open Ended:

## 1] Who are Fetch’s power users?

**Identifying Fetch’s Power Users: The 75th Percentile Approach**

1. **Capturing Highly Engaged Users**
   * Fetch’s user base varies in engagement, with the top 25% (75th percentile) being the most active and integral to the platform.
2. **Key Revenue Contributors**
   * Following the Pareto Principle, this segment drives most transactions and revenue, playing a crucial role in brand partnerships.
3. **Balancing Depth & Scale**
   * The 75th percentile offers a broad yet focused view, avoiding the extremes of an overly exclusive (top 1-5%) or diluted (entire user base) approach.
4. **Ensuring Consistent Activity**
   * Users in this group have at least two transactions, demonstrating sustained engagement with promotions, rewards, and new features.
5. **Alignment with Fetch’s Business Model**
   * These power users provide rich shopping data, strengthen brand relationships, and drive organic growth as early adopters.
6. **Optimized Resource Allocation**
   * Prioritizing these users ensures maximum ROI through targeted marketing, exclusive rewards, and early feature access.
7. **Industry Best Practices**
   * Leading loyalty programs and digital platforms use the top quartile to benchmark engagement and refine strategies.

**Why Not Other Methods?**  
❌ Top 1-5%: Too narrow, missing many valuable users.  
❌ Averages: Distorted by outliers, leading to misleading insights.  
❌ Entire User Base: Dilutes key findings, reducing strategic effectiveness.

**Conclusion**  
The 75th percentile is the optimal threshold for identifying Fetch’s power users, ensuring:  
✔ Revenue growth through high-value engagement.  
✔ Strong brand partnerships with loyal, active customers.  
✔ Efficient resource use via targeted strategies.

## SQL Query:

WITH UserActivity AS (

-- Count transactions per user, considering only users with at least 2 transactions

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

),

PercentileCalc AS (

-- Compute the 75th percentile transaction count dynamically

SELECT

PERCENTILE\_CONT(0.75) WITHIN GROUP (ORDER BY transaction\_count)

OVER () AS seventy\_fifth\_percentile

FROM UserActivity

),

FilteredUsers AS (

-- Select users whose transaction count is greater than or equal to the 75th percentile

SELECT

ua.USER\_ID,

ua.transaction\_count

FROM UserActivity ua

CROSS JOIN PercentileCalc pc

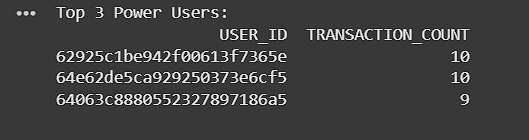
WHERE ua.transaction\_count >= pc.seventy\_fifth\_percentile)

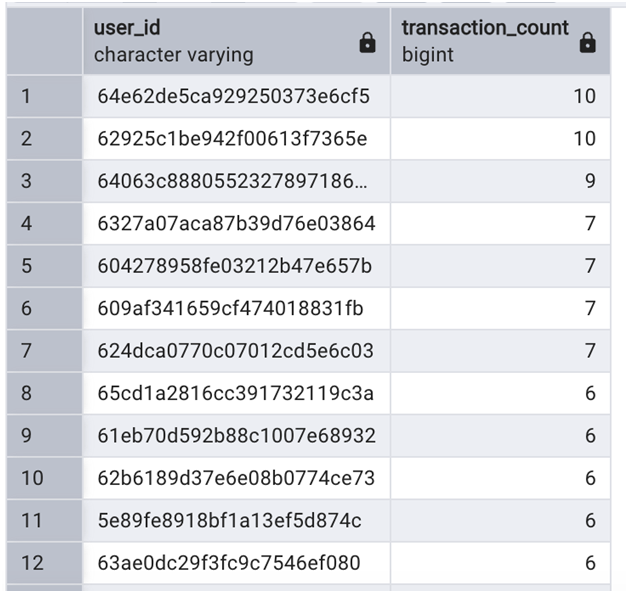
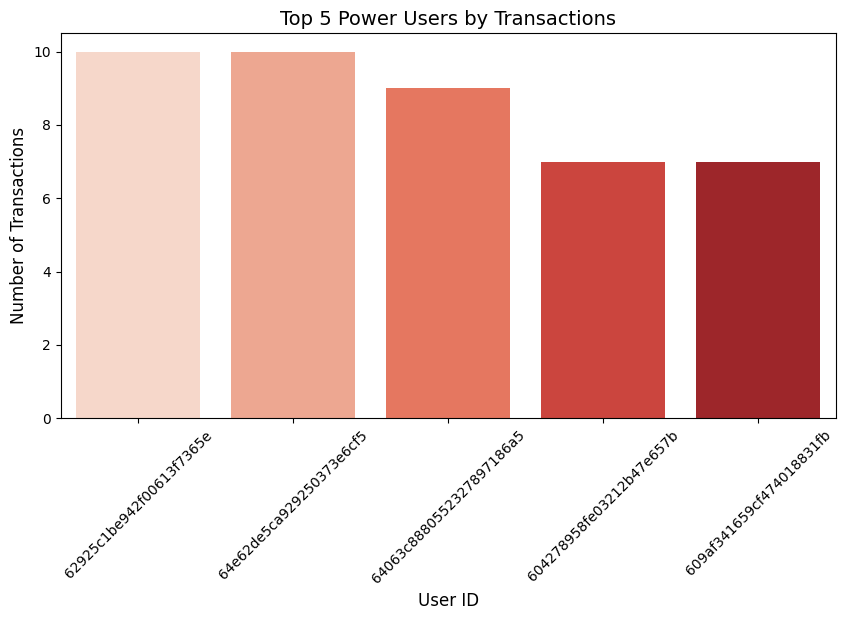
-- Final output: list of power users sorted by transaction count

SELECT \*

FROM FilteredUsers

ORDER BY transaction\_count DESC;





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

#### 1. Total Sales as a Measure of Consumer Preference

* **Accurate Demand Indicator:** Total sales provide a direct measure of consumer spending, making it a reliable gauge of brand popularity.
* **Market Leadership:** The brand with the highest sales in the Salsa category demonstrates strong market presence in both sales volume and revenue.

**2. Relevance to Fetch Rewards**

* **Strengthening Brand Partnerships:** Prioritizing top-selling brands helps Fetch optimize partnerships, ensuring relevance for both users and brands.
* **Enhancing User Engagement:** Showcasing popular brands within the app encourages users to interact more, driving higher participation and transactions.

### ****Business Context: Why Total Sales Matters****

* **Revenue Alignment:** Tracking total sales aligns with Fetch’s goal of understanding consumer spending patterns.
* **Strategic Insights:**
  + **For Users:** Highlighting leading Salsa brands enhances shopping experiences and engagement.
  + **For Brands:** Fetch can leverage sales data to secure better partnerships, offering tailored promotions and incentives.
* **Category Growth:** Higher-selling brands contribute significantly to Fetch’s overall growth, benefiting both the platform and its partner brands.

#### ****Summary:****

Total sales effectively measure brand success and consumer preference in the Salsa category. This metric benefits Fetch by strengthening partnerships, driving user engagement, and supporting business growth. Using RANK over LIMIT ensures a fair and scalable approach to identifying top brands.

## SQL Query:

WITH BrandSales AS (

SELECT

p.brand,

SUM(t.final\_sale) AS total\_sales

FROM transactions t

JOIN products p ON t.product\_id = p.product\_id

WHERE p.category = 'Dips & Salsa'

GROUP BY p.brand)

SELECT

brand,

total\_sales,

RANK() OVER (ORDER BY total\_sales DESC) AS rank FROM BrandSales

ORDER BY total\_sales DESC LIMIT 5;

**3] At what percent has Fetch grown year over year?**

#### ****Understanding the Dataset Limitations****

* The dataset covers **only June to September 2024**, preventing a full year-over-year comparison.
* As a result, the calculated growth percentages for **2024 may not accurately reflect** Fetch’s actual growth trajectory.

#### ****Scalability & Business Value****

* The SQL query is **built to scale**, automatically incorporating future years and dynamically recalculating growth rates.
* Despite incomplete data, it provides a **framework** for consistent growth analysis over time.
* Short-term trends (e.g., **June to September growth**) can still be assessed to gain preliminary insights.

#### ****Why Communicate Data Limitations?****

* **Transparency**: Helps stakeholders interpret results accurately.
* **Future-Readiness**: Demonstrates that Fetch Rewards has a **scalable approach** to track growth as more data becomes available.

#### ****Key Takeaways****

* While the current dataset restricts **full-year growth analysis**, the methodology ensures Fetch is **prepared for long-term tracking**.
* As more data accumulates, this approach will **support comprehensive reporting** and drive **data-informed decision-making**.

## SQL Query:

WITH YearlyTransactions AS (

SELECT

EXTRACT(YEAR FROM TO\_DATE(PURCHASE\_DATE, 'YYYY-MM-DD')) AS transaction\_year,

COUNT(DISTINCT RECEIPT\_ID) AS total\_transactions

FROM prov.transactions

GROUP BY EXTRACT(YEAR FROM TO\_DATE(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; Part 3: Communication with stakeholder

**Subject:** Key Insights & Next Steps for Fetch Rewards

Hey Team,

Wanted to share key findings from my analysis of Fetch’s transaction and product data—highlighting user trends, data gaps, and opportunities for growth.

### ****Key Insights****

* **Top Brands Drive Engagement:** Users 21+ engage most with a few key brands, with the top five making up a big share of receipts—great potential for deeper brand partnerships.
* **Power Users Matter:** The top 25% of users drive most of the activity. Targeted loyalty campaigns could boost retention.
* **Health & Wellness Growth:** Millennials & Gen Z dominate this category—strong opportunity for focused marketing.
* **Salsa Category Standout:** One brand leads in Dips & Salsa sales, signaling potential for category-specific promotions.

### ****Challenges & Next Steps****

* **Data Gaps & Inconsistencies:** Missing product details (CATEGORY\_4, MANUFACTURER, BRAND) and placeholder values need cleanup. Suggesting standardization & automated checks.
* **Limited Growth Analysis:** The dataset only covers June–Sept 2024, so year-over-year trends are incomplete. As more data comes in, this approach will scale seamlessly.
* **Strategic Actions:** Prioritize data enrichment, develop engagement strategies for power users & top brands, and ensure transaction data accuracy.

Let’s align on next steps, happy to set up time to discuss. Let me know what works!

Best,  
Mukesh Jaisankar