Subject: **Key Insights from Fetch Data Analysis**

Dear Stakeholders,

I hope this message finds you well!

I wanted to share the results of my recent analysis of Fetch's transaction data, highlighting an interesting trend and key data quality issues that need attention. Here's a summary:

Key Insight: CVS Dominates Among Long-Term Users
Our transaction data reveals a compelling trend: CVS dominates among users who have had their accounts for at least six months, generating $72 in total sales. This far surpasses the next top brands: DOVE: $30.91, TRIDENT: $23.36.

This significant gap highlights CVS's performance and suggests it is the most preferred brand among long-term Fetch users.

One plausible explanation for CVS's dominance is its strong market presence and alignment with user preferences. CVS is a well-established retail and pharmacy brand, offering a wide range of products that cater to everyday needs. Additionally, CVS may be leveraging frequent promotions, rewards programs, or exclusive deals that resonate particularly well with Fetch's long-term users.

This trend presents a valuable opportunity for Fetch to deepen its partnership with CVS and replicate its success with other brands. Here are a few actionable strategies:

1. Exclusive CVS Promotions: Collaborate with CVS to offer Fetch users exclusive discounts or rewards, further incentivizing engagement.
2. Targeted Campaigns: Launch campaigns that highlight CVS's value proposition to newer users, encouraging them to engage more frequently.
3. Cross-Brand Partnerships: Explore partnerships with other top-performing brands (e.g., DOVE, TRIDENT) to create bundled offers or rewards, boosting their sales and engagement.

By leveraging CVS's success and addressing the engagement gap with other brands, Fetch can optimize user engagement, strengthen brand partnerships, and drive overall growth.

Data Quality Issues and Outstanding Questions
While analyzing the data, I identified several data quality issues across the Product, Transaction, and User tables that need resolution to ensure accurate reporting and analysis:

<u>Product Table</u>

- **Missing Values**:
    - CATEGORY_4: 92% missing, making product categorization incomplete.
    - MANUFACTURER and BRAND: 27% missing, leading to inconsistent records.
    - BARCODE: 0.5% missing, but duplicates exist (2 duplicates), violating uniqueness.
- **Duplicate Rows**: 57 duplicate rows, causing data redundancy.
- **Mixed Data Types**: MANUFACTURER contains both numeric and string values, complicating analysis.

Transaction Table

- **Missing Values**:
    - BARCODE: 11.5% missing, affecting product identification.
    - FINAL_QUANTITY and FINAL_SALE: Missing values and mixed data types (e.g., 'ZERO' instead of 0).
- **Duplicate Rows**:
    - RECEIPT_ID: 25,560 duplicates, undermining its role as a unique identifier.
    - 171 duplicate rows, leading to data redundancy.
- **Date Issues**: 94 rows where SCAN_DATE is before PURCHASE_DATE, creating timeline inconsistencies.

User Table

- **Missing Values**:
    - LANGUAGE: 30.5% missing.
    - GENDER, STATE, and BIRTH_DATE: Significant missing values, though not critical for analysis.
- **Date Anomalies**: 1 record where CREATED_DATE is earlier than BIRTH_DATE, indicating invalid data.
- **Inconsistent Data**: GENDER contains variations like "non_binary" vs. "Non-Binary," requiring standardization.

Outstanding Questions

To address these issues and improve data quality, we need clarity on the following:

- **Duplicate Receipt IDs:** Should we always retain the nonzero FINAL_QUANTITY/FINAL_SALE values, or could zeros sometimes represent legitimate adjustments (e.g., returns or discounts)?
- **Outlier Transactions:** Should wholesale or bulk purchase receipts be included in our analysis, or should they be treated as anomalies?
- **Duplicate BARCODEs in Products:** Could these discrepancies be due to rebranding or data entry errors? What is the best approach to validate the correct product-brand associations?
- **Potential Duplicate Users:** How should we handle cases where multiple accounts might belong to a single individual?

Request for Action

To move forward, I'd appreciate your support with the following:

- **Clarify Data Collection Processes**: Can we connect with the data engineering team to understand why certain fields are missing or inconsistent?
- **Data Cleaning Support**: Can we get help to address duplicates, missing values, and date anomalies?

Your guidance on these questions would be invaluable as we refine our analysis. Please refer to the attached Jupyter Notebook for a more detailed exploration.

Thank you for your time and support. Looking forward to your thoughts!

Best regards,
Vivian Tian