Hi Team,

I hope this message finds you well. I've recently completed a preliminary review of our core data tables, and I would like to share some findings and data quality issues that could potentially impact our analysis and decision-making processes. I have broken down my analysis for each table and have also highlighted any questions or suggestions I might have regarding those which would aid me in a more detailed analysis.

Users Table:

- **Missing Users Data:** When mapping the users table with the transactions table, only 262 out of 50,000 transaction records could be joined. This discrepancy indicates that many user IDs in the transactions table are not present in the users table. **Could you provide the complete dataset for the users table to facilitate comprehensive analysis?**
- The 'language' column has a 30% missing data rate. Missing data in 'gender' (5.89%) and 'state' (4.8%) columns may affect demographic analyses
- The 'birth date' column has about 1.272% of records defaulting to 1970-01-01, indicating placeholder values that may skew age-related metrics. – **Is there a reason why this place holder value is being used?**
- Gender Simplification: The gender column initially had 11 values with redundancies, such as "my gender isn't listed" and "not listed." I've reduced this to 8 categories for clearer analysis. **Can we modify data process to eliminate these redundant gender categories?**

Products Table:

- **Missing Barcode Data:** While attempting to map the products table with the transactions table, only 24,854 out of 50,000 transaction records could be successfully joined. This gap highlights that many barcodes listed in the transactions table are missing from the products table. **Could you supply the complete dataset for the products table to ensure a thorough analysis?**
- Both 'brand' and 'manufacturer' columns are missing approximately 26% of their data, challenging our ability to perform brand-specific analysis. **Can we make meaningful interpolations based on any other information we might have regarding these products using there store name and sale value?**
- Placeholder Values: About 10.28% of the entries in the manufacturer column are labeled as "placeholder manufacturer," and 2.01% of entries in the brand column are labeled as "brand not known." **Do we have any information about these brands and manufacturers, or if they are missing can we leave them as blanks to ensure consistency?**
- A significant 92.02% of entries in the 'CATEGORY_4' column are missing, which limits detailed product categorization. **Can we make effective categorization for this?**

Transactions Table:

- **Duplicate Entries**: Every transaction currently creates two rows per receipt ID, regardless of the number of products sold. **Could we streamline the data collection to generate a single row per receipt when only one product is involved, capturing its quantity and final sale value directly?**

- **Final Quantity Accuracy**: Approximately 0.22% of entries in the 'final quantity' column are non-integer values. **Should these cases be rounded to the nearest integer, or is this variation expected?**
- In some cases, the final quantity is erroneously marked as the string "zero" in what should be a numeric column, further complicating data accuracy
- Out of the 24,795 unique transactions (based on concatenation of receipt id and barcode) 1.31%, either lack a final sale value or have a final sale value set to zero. **Can we understand why this is the case?**
- Barcode Completeness: The barcode column itself has 11.52% missing values, adding complexity to transaction identification and tracking. **Can we understand the reason for missing barcode fields?**

Interesting Trends: There are a few key interesting trends which I could gather from analyzing the YoY percentage growth of new users and have some ideas on it:

Decline in New User Signups There was a 42% YoY decline in new user registrations for 2023, suggesting that Fetch Rewards app may be reaching a saturation point. I have several strategies in mind to potentially boost new user acquisition and look forward to discussing these

1. **Store Preferences**: Walmart dominates transaction volume, accounting for 35.86% of the total sales, followed by Costco (6%), Sam's Club (4.9%), and CVS (4.3%). Targeting other well-represented but underutilized stores with attractive offers could enhance user engagement and acquisition.

2. **Gender Demographics**: Women make up over 65% of our user base. By tailoring our marketing efforts to other demographics, we could increase their signup rates.

3. **Age Groups**: The 20-30 year age group is the most active on Fetch, comprising over 25% of all users. While continuing to engage this core demographic, we could also introduce targeted incentives for users over the age of 50 to expand our reach within this segment.

To address the key data quality issues and refine our understanding of the data, I need the following assistance and information to effectively resolve these outstanding issues:

1. **Data Completeness Enhancement:** Complete data for users and products table to be able to map it to all the transactions present in transactions table. Guidance and support in improving the data completeness for 'language', 'gender', 'brand', and 'manufacturer' columns to bolster our demographic and brand-specific analyses.
2. **Clarification on Data Protocols:** Detailed explanations on the use of placeholder values and potential measures to minimize their presence across datasets.
3. **Data Management Tools:** Implementation of advanced tools for data cleaning and deduplication, particularly to address the issues observed in the 'transactions' table such as duplicate entries and inconsistent data entries.

Thank you for your attention to these matters and please feel free to reach out for further information on any of these points. Thank you so much!

Best Regards - Ashvin Raj (Aspiring Data Analyst – Fetch Rewards)