Fetch Data Analysis Submission

First: explore the data -

Note: this is a text doc, for query exploration and thought process please refer - Query_1_explore_data.pdf

1.) Are there any data quality issues present?

Yes, several data quality issues were identified during the exploration process.

Below are the key findings for the above two points :

1. Data Integrity Issues

- The transactions table contains **non-numeric values** in the final_quantity field (e.g., 'zero'), which had to be converted to 0 for proper numerical analysis.
- The products_staging had **NULL barcode values** (4,025 records), making them unusable as primary keys. These records had to be removed.
- The user_staging had **inconsistent or missing values** for language and gender, making demographic analysis unreliable. Some users have missing birth_date, making it difficult to categorize them into generational segments.

2. Duplicate Records

- **Transactions Table:** Identified **161 duplicate transaction records** removed to prevent inflated sales and transaction counts.
- **Products Table:** Found **185 duplicate barcodes**, which were eliminated before inserting data into the final products table to maintain data accuracy.

Impact:

- These duplicates could have **skewed key business metrics**, such as total revenue, product sales, and customer activity.
- However, considering Fetch Rewards' **business model**, where growth is primarily measured by **user engagement and app activity rather than just sales volume**, the duplicate records in transactions and products may not significantly affect the company's core growth insights.
- The focus should remain on **tracking unique users, receipts scanned,** however, duplicates should be removed.

3. Foreign Keys Integrity Issues

- Only **91 users** in the transactions table exist in the user table, suggesting that many transactions reference **nonexistent users**.
- Only **6,562 barcodes** in the transactions table match the products table, meaning many transactions are linked to **invalid product references**.
- If foreign key constraints were enforced, this would result in **over 18,186 transactions being removed** due to missing user or product references.

2.) Are there any fields that are challenging to understand?

Yes, A few fields stood out as challenging to understand due to their inconsistencies or lack of clear definitions:

User Table

- **language** - The field contains inconsistent values and unclear categorization.
- **gender** - The field contains ambiguous values that don't follow a standardized format.

Transaction Table

- **final_quantity** - Some entries contain non-numeric values (e.g., 'zero' instead of 0), which require manual cleaning. Additionally as discussed above some are missing, making it difficult to determine product quantities in certain transactions.
- **final_sale** - The presence of NULL values in sales transactions raises questions about whether those transactions are valid or need further reconciliation.
- **barcode** - Some transactions reference barcodes that do not exist in the product table, raising concerns about missing product mapping.

Product Table

- **barcode** - There are duplicate barcodes in the product dataset, which should typically be unique identifiers.
- **category_1, category_2, category_3, category_4** - The hierarchical structure of product categories is unclear; even records are missing.