I have completed the analysis of the Users, Transactions and Products datasets in order to get insights about Fetch’s user base and transaction patterns. The Users, Transactions and Products data sets have undergone an initial assessment to reveal information about Fetch user statistics and transaction performance. The data quality problems and significant SQL query results lead to actionable steps for this analysis improvement.

**Data Quality Issues**

Multiple data quality problems affect how the datasets can be used for data analysis. Here’s a detailed breakdown:

1. Missing Values

Users Table:

* BIRTH\_DATE: 3,675 missing values.
* STATE: 4,812 missing values.
* LANGUAGE: 30,508 missing values.
* GENDER: 5,892 missing values.

Transactions Table:

* BARCODE: 5,762 missing values.

Products Table:

* CATEGORY\_1: 111 missing values.
* CATEGORY\_2: 1,424 missing values.
* CATEGORY\_3: 60,566 missing values.
* CATEGORY\_4: 778,093 missing values.
* MANUFACTURER: 226,474 missing values.
* BRAND: 226,472 missing values.
* BARCODE: 4,025 missing values.

The implications for this are:

* The absence of BIRTH\_DATE information makes it impossible to perform age-based queries since user age calculations become inaccurate specifically for users aged 21 and older.
* The absence of BARCODE in Transactions and Products prevents the successful linkage of transactions to particular products.
* The high number of missing values in both CATEGORY\_3 and CATEGORY\_4 indicates that these fields lack validity for precise product classification.

1. Duplicates

* Users Table: No duplicate IDs (0 duplicates).
* Products Table: 4,209 duplicate BARCODEs.
* Transactions Table: 25,205 duplicate RECEIPT\_ID + BARCODE combinations.

The implications for this are:

* The presence of duplicate BARCODE values in the Products table leads to problems with joining or aggregating data correctly.
* The combination of RECEIPT\_ID + BARCODE in Transactions may represent either valid product duplicate entries or data entry mistakes.

1. Data Types

Both Transactions and Products store their BARCODE values as float64 data type while most fields exist as object (strings).

The conversion process transformed FINAL\_QUANTITY and FINAL\_SALE from object format to float64 data type during the cleaning phase.

1. Invalid Values

Transactions Table:

* The cleaning process replaced all alphabetic characters in 12,500 FINAL\_QUANTITY rows with numerical values of zero.
* The cleaning process did not reveal any negative numbers in FINAL\_QUANTITY or FINAL\_SALE.
* The implication for this is that the alphabetic entries appearing in the FINAL\_QUANTITY database require attention because they point to inconsistent data entry that needed correction.

1. Referential Integrity

* The transaction data contains 49,738 entries with USER\_ID values that do not exist in the Users table.
* 19308 transactions lack BARCODES in the Products database. This leads to user and product data linkage failing for many transactions, which means analysts cannot analyze those cases.

**Findings from SQL Queries**

The following information emerged from running SQL queries against an in-memory SQLite database system.

1. Top 5 Brands by Receipts Scanned (Users 21+)

The query produced a list of brands that users who reached age 21 scanned through their receipts.

* NERDS CANDY: 3 receipts.
* DOVE : 3 receipts.
* TRIDENT: 2 receipts.
* SOUR PATCH KIDS: 2 receipts

1. Top 5 Brands by Sales (Accounts 6+ Months Old)

This analysis determined the total sales from brands which users maintained for at least 6 months before the most recent scan date.

* CVS: $72.00
* TRIDENT: $46.72
* DOVE: $42.88
* COORS LIGHT: $34.96
* AXE: $15.98

The data indicates long-term CVS customers make the most purchases while personal care products and beverages follow in a separate category as observed through sales figures.

1. Power Users

The query revealed that 10 users attained the highest positions based on the distinct receipt counts they scanned.

* Top user (064e62de5ca929250373e6cf5): 10 receipts.
* Second user (62925c1be942f00613f7365e): 10 receipts.
* Third user (64063c8880552327897186a5): 9 receipts.
* Others: 6-7 receipts each.

A limited group of very involved users who checked 6 to 10 receipts demonstrates their potential role as ideal candidates for target marketing initiatives.

An interesting trend noted was users who scan more than ten receipts belong to an especially active segment according to the receipt data. Currently active users who scan 10 receipts per month might serve as potential candidates to join loyalty programs while providing feedback about Fetch.

Outstanding Questions and Next Steps

The analysis requires additional feedback regarding specific recommendations to resolve the discovered problems.

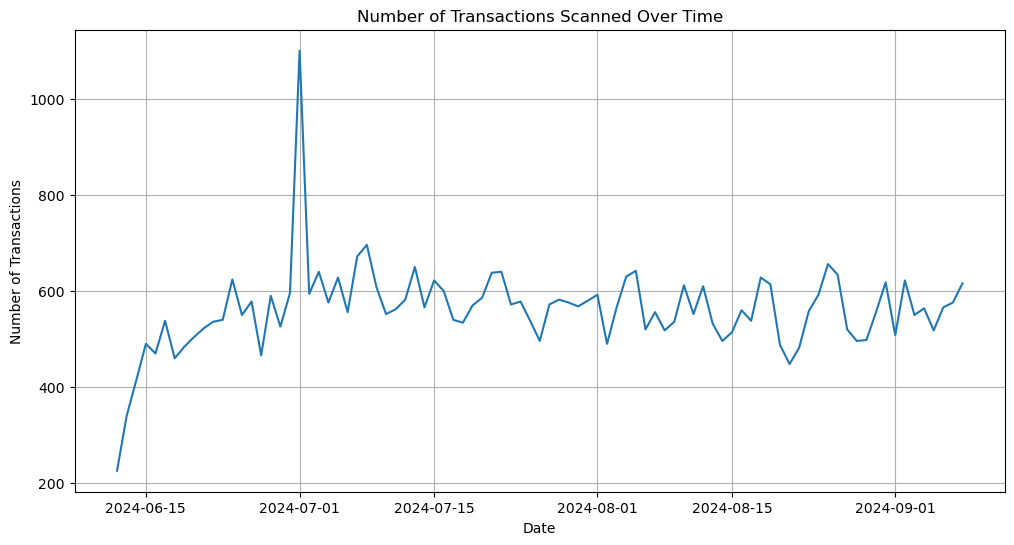
**Visualizations**

The analysis consists of two visualizations that enable one to have a better understanding of features in the dataset.

1. **Transactions Over Time**

The analysis displays daily transaction volumes through a line chart which uses SCAN\_DATE as its basis.

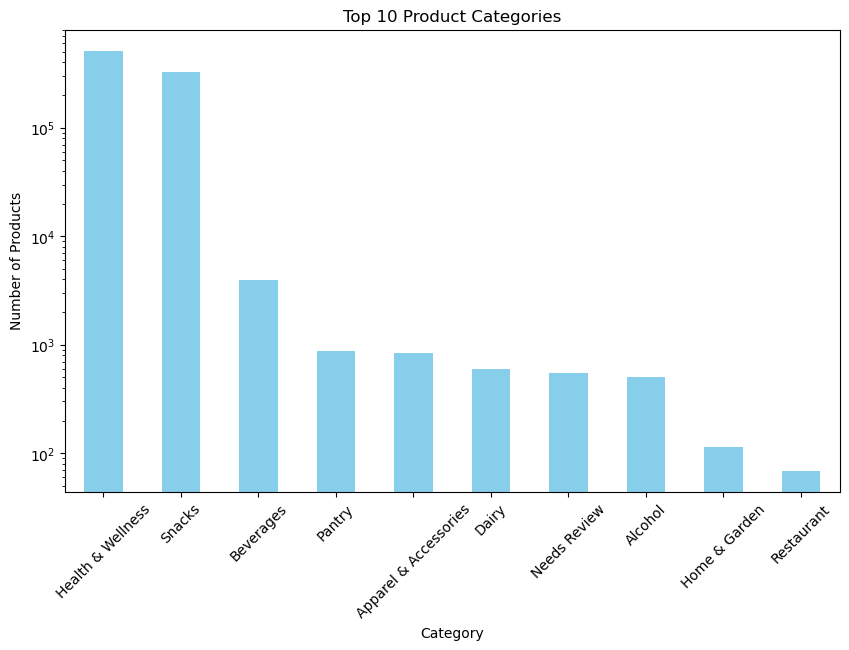
The time frame of the plot extends from June 12 through September 8 in 2024, but the specific peak dates remain undisclosed in the incomplete output.



1. **Top 10 Product Categories**

The bar chart below presents the ten most common CATEGORY\_1 values through a logarithmic scale because the data shows wide variations (representing 845,552 products).

The analysis revealed 27 different categories among the datasets, yet the top ten categories are plotted. The analysis shows essential product categories within Fetch's business operations, such as Health and wellness, as well as snacks, are more dominant.



**Data Rules Clarification**

The system expects to encounter duplicate BARCODEs in the Products section when different product variants exist in the system. What measures should be taken when duplicate BARCODEs occur in the Products table?

The procedure for Transactions with missing BARCODEs should include imputation, exclusion, or marker-flagging.

**Metadata Request**

Please outline the structure of product categories between CATEGORY\_1 to CATEGORY\_4 and show their intended purposes.

**Definition Confirmation**

The definition of power users requires clarification through a specific metric that quantifies their status. Should the metric consider receipt count or total sales volume? This will guide future analyses.

Your feedback is eagerly awaited while I express gratitude for your helpfulness.