Stakeholder,

Here are the results of my recent investigation of our product, user, and transaction files. To summarize my findings, there are some data quality issues preventing me from getting useful results. I have written some database queries that will provide the results we need quickly once those data quality issues are resolved.

**Data quality issues**

There are three major data quality issues with the existing data which need to be addressed.

1. Many of the records are incomplete. Approximately 27% of the records in the products file do not have a manufacturer or brand listed. About 1% of the product records and 11% of the transaction records do not have a barcode. The barcode is used to cross-reference product records and transaction records, so these missing values prevent transactions and products from being properly linked.

There are also other fields where a large number of records are missing values. In the products file, some of the category fields are blank. In the users file, some users do not have a birth date, language, geographic location, or gender listed. Some users also seemed to enter false birth dates such as 1/1/1901. There is likely no way to change this; the category fields in question are meant to be blank, and many users may decline to enter some personal information or choose to enter a false birth date.

2. There are values with invalid data types in the transaction file. Exactly half of the records in the transactions have non-numeric values for the FINAL\_QUANTITY and FINAL\_SALE fields, both of which should only contain numeric values. For example, some contain the word “zero” in place of the number zero. I wrote a Python script to clean the data set before importing it into SQL. I also noticed that these rows appear to be partial duplicates of other, complete records, so I have been checking for duplicates when writing SQL queries.

3. The transaction file and the users file both appear to be missing large numbers of records. All transaction records in the transaction file contain a user ID, but fewer than 1% of those user IDs also appear in the users file. I also suspect that the transaction file is missing records, since every transaction included in the file took place between June 12, 2024 and September 8, 2024.

Because the transaction and user files only include small subsets of Fetch’s transactions and users, we cannot trust the existing results from my queries. With more complete data sets, I should be able to run the same queries to get much more useful results.

**Interesting trend in the data**

After querying the transaction and user data to determine what percentage of purchases in the Health & Wellness category were made by each generation, I ran a modified version of the same query which covered all categories. There is only one item, with a total sale price of $1.97, purchased by a member of Generation Z in our file. 26 percent of the users in the user file belong to Generation Z, so this discrepancy is worth investigating further.

**Request for action**

To make sense of the conclusions, the single most important things that I need are complete user and transaction data sets. With the partial files I have, the sample size is too small to gain useful insights about the purchases made by users.

In addition to these data sets, I would appreciate a review of the null values in the manufacturer and brand fields in the products data set due to the large portion of missing data. Additionally, we should search for the barcode data for any product in our database which is missing a barcode.

Let me know if you have any further questions!

Thanks,

Andrew