# Data Quality Issues:

**Transactions Table:**

* 5,762 rows are missing BARCODE values, and 12,765 rows have a null value in FINAL\_SALE.
* FINAL\_QUANTITY contains zero values, which are then converted to 0.
* 335 duplicate values exist in the transaction data.
* Rows with FINAL\_SALE or FINAL\_QUANTITY as 0 are repeated, but corresponding rows contain the correct FINAL\_SALE and FINAL\_QUANTITY values.
* 47 receipts show a scan date earlier than the purchase date.
* BARCODE contains invalid values, such as -1.

**USERS Table:**

* Missing data includes 3.7% of BIRTH\_DATE, 4.8% of STATE, 30.5% of STATE, and 5.9% of GENDER values.
* Inconsistent formatting is found in the GENDER column, with phrases like “Prefer not to say,” “Non-binary,” and “Gender isn’t listed” written in various ways.

**PRODUCTS Table:**

* 4,025 products lack a barcode, with 92% of CATEGORY\_4, 27% of MANUFACTURER, and 27% of BRAND column values missing.
* Duplicate barcodes exist, representing different brands.A screenshot of a computer

  Description automatically generated

# Data Cleaning:

**Transactions Table:**

* Converted the BARCODE column to a string to preserve leading zeros removed during CSV import due to INT format.
* Replaced null values in BARCODE with “NA.”
* Converted FINAL\_QUANTITY zeros and null values in FINAL\_SALE to 0, changing both columns to FLOAT type.
* Removed the 335 duplicate values.
* Changed PURCHASE\_DATE and SCAN\_DATE to DateTime format.
* There are some rows with the same RECEIPT\_ID and BARCODE, where either FINAL\_SALE or FINAL\_QUANTITY is zero. In such cases, there are duplicate rows with the same RECEIPT\_ID, BARCODE, STORE\_NAME, and USER\_ID, where the other row contains the correct values. So, I removed the rows containing either FINAL\_QUANTITY or FINAL\_SALE is 0.

**USERS Table:**

* Converted CREATED\_DATE and BIRTH\_DATE to DateTime format.
* Replaced null values in GENDER, LANGUAGE, and STATE columns with “NA.”
* Replaced null values in BIRTH\_DATE with random values from existing records to maintain distribution.
* Standardized the GENDER column values by mapping terms to “Prefer not to say,” “Non-Binary,” and “My gender isn’t listed.”
* Created an AGE column by calculating the difference between BIRTH\_DATE and the current date for each user.

**PRODUCTS Table:**

* Removed rows without a barcode, as they are not useful for analysis.
* Retained only the first instance of each duplicated barcode, removing subsequent duplicates.
* Dropped the CATEGORY\_4 column due to 92% missing values.
* Replaced remaining null values with “NA.

# Closed-ended Questions:

1. **What are the top 5 brands by receipts scanned among users 21 and over?**

A screenshot of a computer code

Description automatically generated

1. **What are the top 5 brands by sales among users that have had their account for at least six months?**

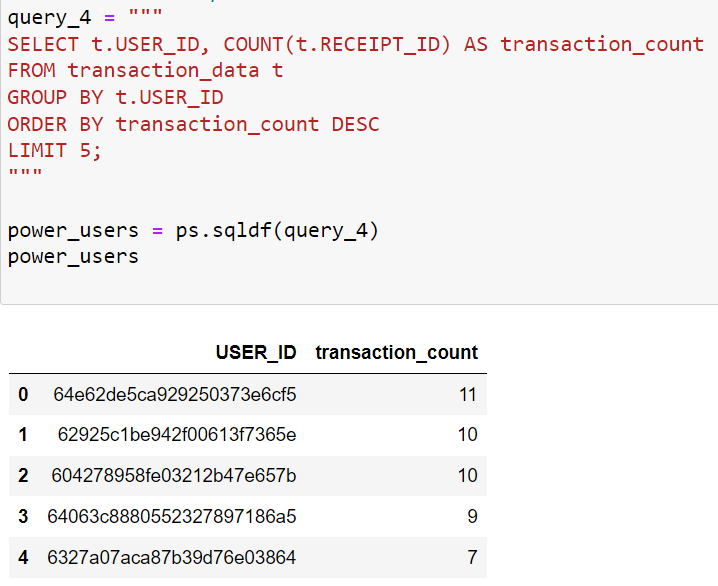
A screenshot of a computer

Description automatically generated

# Open-ended questions:

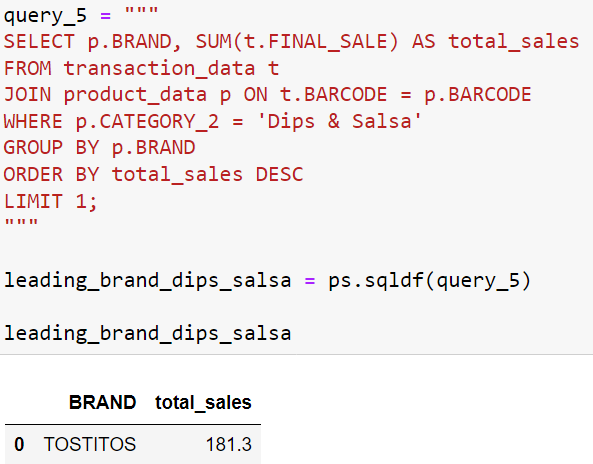
1. **Who are Fetch’s power users?**

Fetch's power users are customers who exhibit high engagement and frequent activity on the platform. So, I took Fetch Power users as the users with the highest receipt count.

****

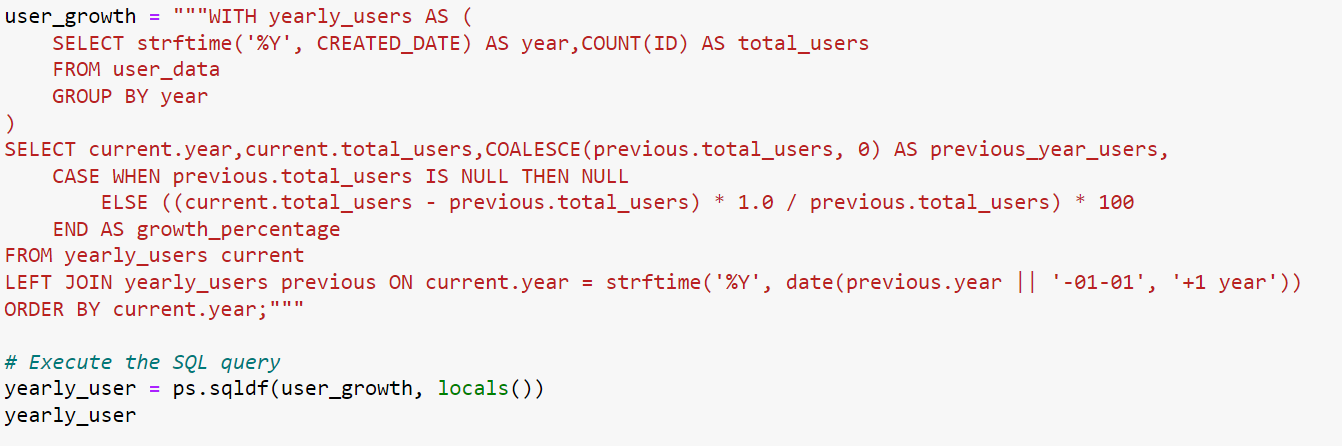
1. **Which is the leading brand in the Dips & Salsa category?**

Here, the leading brand is taken as the brand with the highest FINAL\_SALE. So, for this data is taken from the transactions table and the products table. Joined both tables on BARCODE column and took the CATEGORY\_2 as ‘Dips & Salsa’ grouped by brand and then ordered by sum of FINAL\_SALE and then took the top product. I have taken the cleaned data for this.



1. **At what percent has Fetch grown year over year?**

For this, I considered the growth in terms as the number of users created over time. We can take sales growth also, but we have only 3 months of data. I assumed created date in user\_data table as user created date and the percentage change in current year users to previous year users is calculated as growth percentage.



A table with numbers and text

Description automatically generated

# Challenging fields to understand:

1. **Transactions Table:**

* The 335 duplicate values: Are they genuinely duplicated transactions, or could they represent cases where the user bought two items with each scanned twice? Should we treat them as duplicates?
* 47 receipts with SCAN\_DATE earlier than PURCHASE\_DATE: How should these be interpreted? Are they potentially fraudulent transactions, and should they be removed?
* Rows with FINAL\_SALE or FINAL\_QUANTITY as 0 but with corresponding rows containing correct values: Could these indicate returned items or data entry errors?
* Items with BARCODE as -1: What does this represent, and how can a barcode have a value of -1?

1. **Users Table:**

* BIRTH\_DATE column has missing values and many records have same birth date 1970-01-01. Is this the default date? Should we replace null values with the default date.

1. **Products table:**

* Two brands share the same barcode: Which brand should retain the barcode, and which should be removed?
* Over 4,000 items lack a barcode: What might be the cause? Should we remove these rows?
* The CATEGORY\_4 column has 92% missing values: Is this column necessary for analysis?