Hello Team,

Here are some findings from our consumer data.

First, data questions.

We have key data fields in product that are missing. The main issue being missing barcode data, this makes it impossible to link that product to a purchase. Let’s ensure we get all barcode information as a required data field when inputting into our system. We also have some products that have a barcode, but don’t have any category/brand information tied to them. This again makes it difficult to perform insights.

Transaction data was a bit messy, with duplicative rows with some having sales data and some not. I am guessing this is a side effect of how that data is gathered from the source transaction system. Lets invest time into our ETL process to run a job that cleans that data up for us.

Insights:

User user\_id 6682b24786cc41b000ce5e77 has been our top engaged customer, with an average of 1 transaction per 107 days. Lets explore that users behavior to try to understand why they are more engaged with the platform than others. Are there marketing notifications they are engaging with, markets/brands, etc?

Our YoY growth looks like this:



Our peak user creation year was 2022 with 26,807 account creations, with an early exponential increase from 2016 – 2020 and a more mature growth from 2021-2022. 2023-2024 we have declined in account creation. However, account creation is only one factor. We need to determine active users, stale accounts, and lifetime value for our active users. Account creation growth is not indicative of quality of users.

I have a query in Git that would answer this question:

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

However, the IDE I used for this assignment is not the best, and it just spun on my subquery, apologies.