Respected Sir/Ma’am,

I wanted to highlight few concerns about our current dataset. The three datasets have various issues which might cause major roadblocks while analyzing our customers, brands behavior. Based on our business model we value the brands associated with Fetch rewards immensely through the partnerships and providing accurate data to these brands is of utmost criticality to honor these partnerships.

Following are the major concerns with our current data:

* **Data Inconsistency:** Brands database isn’t consistent with our receipts database. Many of the receipts our customers are scanning have various brands and their barcodes, but these barcodes are missing in the Brands database which creates anomalies while understanding the top brands for the recent months. There may be some brands which could leverage our services, but we cannot share them the exact details about how frequently people are buying these brands.
* **Ambiguous Data**: There are many receipts without user ids which is one of the major identification factors for any Receipt. Without user id that receipt won’t help us understand the customer behavior and wouldn’t provide credibility to consider the brand’s impact without any user details available.
* **Inaccurate Data**: In some receipts there is a discrepancy between among the total number of items bought by a user. Either there was a data loss while capturing the item details or the total items bought by a user is captured inaccurately. Similarly, Final price for a rewards items list doesn’t match the item price and quantity value.
* **Data Duplication:** More than half of the users in the user database are not unique and thus the users list isn’t providing an accurate estimation of total number of active users with their status and signup source data.

My team discovered many such concerns, and these could be resolved easily to better analyze our customer & Brands behaviors. We carried out multiple quality checks in Python & SQL to arrive at these conclusions. One key action item would be making our datasets consistent and running a pipeline which updates all our datasets at regular intervals. For avoiding such concerns, we could also ensure creating smaller datasets with individual receipt items to comprehensively capture all the data related to every item a user has purchased. This would also suffice in capturing their patterns in shopping and buying capacity over a time using a Time series model which could be beneficial in providing a credit line to our valuable customers.

Addressing these issues would be crucial in involving more and more brands with our company and reducing any shortcomings for our exists brands. We could share monthly or weekly dashboards with the brands based on our multiple KPIs and thus venturing into several other domains.

Please let me know if you need any further information. We can setup a meeting to discuss all such concerns and my team has already some insightful suggestions ready for tackling these challenges.

Thanks & Regards

Rajas Kapure.