Part III: Communicate with Stakeholders

Hi @fetch-product-leader:

I've investigated the 3 datasets that were shared. Summarizing the data quality:

1. USER.csv: The gives us info on Fetch's userbase centered around a user id. It appears user submitted and as such is sparse, inconsistent, and inaccurate. State, language, gender, and birth date all have missing values. Gender has 11 distinct values, some of which code for the same thing like "not_listed" and "My gender isn't listed". Birth date has unrealistic values, like those in the year 1900. There is a user with a birth date after the created date.
2. TRANSACTION.csv: This is the core of our database, consisting of the actual receipt information from our users. The main issue here is that the quantity and sale amount columns appear corrupted with 50% of the dataset having either a "zero" quantity or a blank dollar amount. How exactly is this data collected? If it's straight from a computer vision OCR model I think we should incorporate a human-in-the-loop process to validate the data and work to fix this. Could it be due to new variations in receipt structure? We also have ~300 duplicate rows here, which may be related.
3. PRODUCTS.csv: This is our product catalog data. It has 400 duplicate rows. Some other quality issues include inconsistent capitalization between columns, NA/negative barcodes, progressively more missing data the more specific the product category becomes, and placeholder data. Manufacturer and brand also do not have full coverage. Finally, a few dozen barcodes have multiple entries that are not duplicates, meaning that a single barcode points to distinct products. This could be due to a simple category change or a complex product rebrand. I think it would be worth it to start curating this product backlog so we have a clean table to join into our master transactions data.

Fixing these issues is crucial to measuring key vitals of our business. Let's meet at your earliest convenience and we can go over:

1. Transaction data collection process / expected data format from receipts
2. Developing a standardized process for users to enter more accurate demographic data
3. Product catalog curation

Data quality aside, I'm excited to see we've seen double digit growth through the duration of the data set. Our average YoY growth has been:

- **76%** increase in users
- **67%** increase in number of receipts scanned
- **67%** increase in number of unique products scanned
- **80%** increase in total sale amount.

Sincerely,

Jonah Somers

PS: Another large issue is that few users or products from the transactions are covered in their respective datasets, meaning we can't understand a decent portion of user activity at the user level. We need more data on our products and users, so if we mandate some of these collection policies it would certainly help our analytics.