Subject: Data Quality Findings and next steps for Receipts Analysis

Hi [insert business leaders names],

I am reaching out as I have completed analysis for the receipts, brands and user data sets we recently received. I wanted to share some of my findings and challenges as these are crucial to getting accurate reporting and making sound business decisions. The findings are as follows:

Missing Data:

* Approximately 5% (435/7381) of receipts have missing totalSpent values
  + This could be due to fraudulent receipts, bad scanning as some receipts have no barcode data or purchasedItemCounts
  + The vast majority of these empty receipts are coming from a few users so we may have to flag those users accounts.

Duplicate records:

* We found around 1119 duplicate records based on matching \_id values.
  + This could be due to numerous factors including users scanning multiple times, users buying more of a certain item, etc.

Unknown commonalities

* Some of the relationships such as the commonalities between brands and receipts were unable to be found/merged because they had attributes that did not line up with each other.
  + Possibly we need to take in more attributes so they can be easier combined.
  + For example, we look for receipts with the ‘ACCEPTED’ status but the categories currently in place for status are ‘Finsihed’, ‘Submitted’, ‘Rejected’, ‘Pending’ and ‘Flagged’.

We identified these issues by leveraging SQL queries and python based data analytics techniques against the data to validate relationships between receipts, users, and brands.

Moving forward, I believe we should take in more attributes such as the aforementioned problem with brands and receipts. Flagging users who have submitted multiple fraudulent receipts and ensuring our brand data is up to date could also improve our overall data quality.

I would be more than happy to schedule a meeting later on this week or next to further discuss how to optimize our data quality and how to get ready to scale this into production. As we know inconsistent data entries can impact our downstream analytics. Please let me know your availability and I’ll put something on our calendars.

Thank you for your time.

Best regards,

Thay