Hello everyone,

Hope you ‘all doing well!

Firstly, I would like to thank you, for giving me the opportunity to work on fetch rewards data, specifically on receipts, users, and brands data. I am writing this email to outline how I processed data, found possible issues, and my findings. Moreover, to import, clean, and format the data I worked with Python, SQL, and Excel. I have structured the data and performed operations for consistency across all tables, and would like to provide my insights on the given questions:

1. When I was structuring the receipts data for data modeling, I found receipts have an attribute called receipts Item, which in turn has 18 more attributes within it. I would like to understand why receipts item is placed in receipts. Instead of better normalization and quicker access to receipts Item, I split it into a new table.

2. During the process of finding data quality issues, I figured the following problems, could be handled and will improve data assets:

   a) Initially, when reviewed the tables, I checked for receipts and their total spending. I noticed that 39% total spent data was missing for receipts that we had. I believe I am working with only a sample of data, but this number (missing total spend %) should go down across all the data we have for receipts (as it is one of the primary KPIs for our business).
   b) Moreover, I assume our servers might track date-time attributes in UTC format (category of the date/time format), I believe storing this attribute in an understandable format may reduce errors.
   c) Talking about users, after doing some data wrangling, I found out that there were 24% duplicate users. I believe an investigation in this matter is needed and would like to create a plan to reduce duplicate users.
   d) Lastly, I compared barcodes and brand codes between Brands and receipts Item tables (table I created out of receipts), I found out barcodes in each table are unique and brand codes in receipts Item table have more than 90% of the values missing compared to brands table.
   e) While analyzing barcode and brand codes, the foreign key in these tables is a string and I would recommend it to be a number, for better accuracy and consistency.

All the data processing, findings, and data quality issues that I have discussed above are saved in a python notebook. I have attached the notebook below with this email if anyone’s interested to go through it.

Also, I am open to schedule a time to meet, to discuss my findings in detail.

Thank you for your time.

Best Regards,

Viraj Shukla