Subject: Observations about the Receipts Data

------------------------------------------------------------------------------------------------------------------------------------------

Good morning *Product\_Leader\_Name*,

Here are the answers to your questions about the data provided for this first project.

**---What questions do you have about the data?**

From which system/source was the data was exported from?

Why are the tables stored in API format?

How often is this data updated? (Depending on this, some of the query formulas may vary)

How often do I need to make reports with this data?

**---How did you discover the data quality issues?**

At a first glance, I used an API opener tool to see the information with a better indentation (spacing between chunks of text). After that I could see that this API has an Array shape, and it has a List OBJECT in the field "rewardsReceiptItemList".

Also, I was going to use Python to open this data as a data frame and it could not open the file right away because of the formatting issues.

This format is not compatible with a database format that connects standardized tables. Without doing the appropriate cleaning and formatting, this data cannot be used in a database, and we cannot make Queries with it.

**---What do you need to know to resolve the data quality issues?**

Fixing and setting up these data files depend on more tables that are not included in the original group of files. If I could see how all this data is captured and set up that would provide me valuable insight about how to clean and modify these tables before making Queries and building reports with them.

**---What other information would you need to help you optimize the data assets you're trying to create?**

Given the hints and context that the tables contain: We are first talking about some "user"s, "brand"s, and "receipts". Besides that, we need tables information about "Item"s (that belong to the Brands), and "reward"s just to find out which Rewards are available.

It would be crucial to separate the information from the source. The table “receipts” has combined information about “RewardsReceiptItemList”, which should be put apart in two different Data files so it becomes readable for the Queries software, and we could get insights from it.

**---What performance and scaling concerns do you anticipate in production and how do you plan to address them?**

The current format of API arrays doesn't allow us to open the files to start working. Instead, we need to do some cleaning steps that can be avoided. Specially because a functional database that involve the three tables provided still require some other tables. That issues was described in the previous question.

It should be possible to export this data in a Query readable format such as .csv (comma separated values) In case it is not I can still transform this data into a preferred format.

Juan-Diego Ospina

Data Analyst

Fetch Rewards

**\*Fetch rewards logo\***