**Stakeholder Communication**

**Hello,**

**I hope you all are doing well.**

**I have just worked with the input data provided. Since the data was in zipped json files I used python to read, clean and structure the data to a more user-friendly format. Once the data was clean, I ran some sanity checks for quality and following are my observations**

**Nested JSONs were found in both the Receipts and Brands file. Parsing through nested data can lead to an increase in error, and it also slows down the execution of the code. If possible, we can flatten it early on as a pre-processing step for easier further processing.**

**I noticed that the 'brandCodes' are not numerical and my recommendation having worked with databases would be to have the foreign keys numerical to prevent small differences in the string creating problems when joining the data.**

**A sizeable fraction of data is missing for certain fields like finishedDate where 49% of the receipts we do not know when they become invalid and pointsEarned where 45% of the values for the 'pointsEarned' field are missing.**

**There are a lot of (more than half) duplicate records in the Users data. I strongly suggest going over out database, to eliminate duplicate records**

**I found the date formats to be inconsistent. I would recommend setting a consistent date format.**

**A lot of the data is stored in embedded dictionaries (such as date scanned) that require an extra step of data wrangling. It might reduce the margin of error if we remove this but I would love to understand if this is required in some way in the product implementation.**

**Finally,**

**I would like to ask the following questions:**

**How is the data entered for Receipts, Users, and Brands?**

**How are the associated JSON files created?**

**How should I handle empty values? Would you like me to remove those rows with empty values, or should I replace them with fillers?**

**Which columns can we remove from the Receipts table?**

**Please find attached the ER diagram for the dataset mapping. Also attached are few screenshots of the data quality checks.**

**Please find attached a PDF file containing SQL queries that answer few questions.**

**We can set up some time to connect over this and discuss the observations. Feel free to CC anyone who is interested or who can add input to this thread.**

**Thanks and Regards,**

**Roni**