Hello everyone.

I hope everything is going well. I have just processed the fetch rewards data, specifically the receipt, user and brand data, and would like to report some of my findings, possible problems and corresponding solutions.

First, I'll give a little background on how I'm working with this data. Since this data was unformatted messy json files, I used Python and SQL to read, clean and structure the data into a more user-friendly format.Once the data was cleaned up, I performed some data quality checks and had some feedback and questions.

1. As the primary key in the tables, I first went to check if the ids in the three tables were unique ID users, but during the data cleanup I noticed that for the same user, there were multiple records for both the last login time and the creation time. I hope to know the reason why this situation exists

2. The inconsistency in the data exists, which means that there are many receipts in the data with brand codes that do not exist in the brand data. This may have an impact on the data analysis when we aggregate by brand to produce analysis results.

3. I was looking at the statistics of the user's table and noticed that all four columns extracted by the user have large outliers, and wanted to ask about the reasons for this.

4. I noticed that 'brandCode' is not float64 or int64 but object, I think in database work, foreign keys should be numeric to prevent small errors when establishing relationships.

5. A lot of data is stored in embedded dictionaries (e.g. dateScanned) that require additional data processing steps. If we extract the key, we need to go ahead and check if the key value is unique, otherwise to normalize the whole dataset, and I would like to understand if there is a special usage for this format.

I would like to discuss these issues further and schedule time to elaborate on these comments. Looking forward to hearing more from you.

Bests,

Joshua