Dear Shareholders,

This email is to deliver to you our current project on three tables of datasets, which are separately related to users, brands, and receipts. The following content will mention the questions we have about the data, the data quality issues we discovered, the methods to resolve the data quality issues, new information we need to optimize the data assets, and performance and scaling concerns we anticipate in production and how we plan to address them.

After having an overview of those tables, the **questions we have** are:

1. In the brands table, we are confused about what the criteria are to determine whether a brand is a top brand or not. For example, some brands like Starbucks, and Pepsi, are not listed in top brands.
2. In the brands table, we are confused that some brand names are “test brand @(a list of numbers)”.
3. In the brands table, the category column already shows specific information, but category code shows a lot of missing data. After checking, it was found that the category and the following specific items matched correctly, but the category code was not displayed successfully.
4. In the users table, we are confused about what criteria are used to determine users’ states.
5. In the users table, we are worried about our system operation because some customers’ last login dates are missing.
6. The barcode data in brands table and receipts table seems not the same.
7. In the receipts table, we are confused about what criteria are used to determine whether a receipt should be flagged or rejected.

Secondly, **data quality issues** we discovered:

1. The data is in JSON format files at the beginning. When we try to directly import files into SQL, a programming tool, we found it doesn’t work.
2. And, the reason for it is because data are unstructured (messy), so the first step we did is to clean the data(make it easier to read and analyze). For example, we only kept variables useful for our projects. We changed UNIX epoch time (machine-readable number, 1652925826) to human-readable dates (May 19, 2022 2:03:46 AM GMT). Also, we renamed variables to make them easily readable. Besides, when importing datasets, SQL automatically deleted rows of receipts table with missing data, and since those are pending receipts and do not affect analysis, we don’t need to worry and consider about it.
3. We found there are a lot of data missing in those three tables which somehow affects the result of the analysis except that the case that we mentioned in the second issue about missing data in pending receipts.
4. We analyzed all the variables in brands table, but unfortunately, we didn’t find any foreign keys (links) with the other two tables. Therefore the data in different tables cannot be joined together.

5. We found outliers of brands based on spending of a recepit, but there seems to have no relation with the status of a receipt (flagged, rejected, approved).

What do we need to know to **resolve the data quality issues**:

1. We need to know whether the company can unify the system and database Settings so that the exported data can be processed or entered in the format required by departments, which can improve the allocation of resources and time and improve work efficiency.
2. Also, we need to know whether the database system can have a data entry restriction. For example, this entry box must be filled before submitting the order, or there can be an option box to choose brand names.

What others need to help you **optimize the data assets**:

1. We can subdivide the item list in receipts table into brands, series and capacity lists, instead of directly integrating all the detailed item information into one column. Subdividing the item list will help us identify and classify items more quickly.
2. Find the cause of unnecessary missing values and resolve them. For example, the category merchants that have been identified are not matched category codes. This may be a problem with automatic matching, which should be resolved after fine-tuning. The time when a user creates an account and the last login time are important basis conditions for us to promote. Generally speaking, these two times are easy to record and keep for a long time. The missing value may be caused by the shutdown and maintenance of the database server, resulting in the loss of this part of the data. We should try to back up complete data before each migration and maintenance.

Performance and scaling concerns **anticipation & solutions**:

1. We anticipate there will be a problem and not comprehensive information about brands if we continue to have no foreign key of brands table. The solution we suggest is to add a unique identifier like brand id to receipts table, and then we can connect brands data to receipt data. Also, receipts table already has a bridge with users table, so these three tables will be connected successfully.

Should you need any further information, please do not hesitate to contact us. In the end, we would like to present our heartfelt gratitude towards the shareholders as well the employees for their continued support as well as confidence upon us.

Yours Sincerely,

Data Analytics Team

Fetch Rewards