Hello,

I hope this message finds you well. I wanted to inform you about our recent analysis of the 3 JSON files – receipts, brands, and user.

Major Data Quality Issues: Missing Data, Duplicates, and Formatting (mainly date)

1. Columns like bonusPointsEarned, bonusPointsEarnedReason, finished Date, pointsAwardedDate, purchaseDate, purchasedItemCount, rewardsReceiptsItemsList, totalSpent have missing values.

2. The JSON file has nested structure, and the format of date is not appropriate. It needs to be cleaned to make the better use of this data.

3. When looking for outliers in the receipts.json file we see there are many outliers but cannot say with confidence if we need to disregard those values or not. We need to further investigate.

4. The brands.json file also contains missing value for category, categoryCode, topBrand, and brandCode. We again need to investigate this to see why the data is missing and need to decide what steps to take to counter them.

5. We see that Baking, Beer Wine Spirits, Snacks, Candy & Sweets, and Beverages are the top category.

6. Users file has relatively fewer missing data, but it also has missing data for signupSource, lastLogin, and State.

We need to investigate and check why are there substantial number of missing data as it will cause problems while doing analysis. We might have to have a conversation with the dev team to check if there is any problem in the app or some error while creating the file. There are also instances of duplicate data and we need to check if there are truly duplicate or not.

We need to impute the missing data, get rid of the duplicate data, and clean the data to have standardized format of date into MM-DD-YYYY.

For performance and scaling, we need to ensure that our analysis data pipeline can handle the data effectively.

Let me know when you are available, and we can schedule a call to discuss the issues move ahead with our analysis.

Thank you,

Vidit Gandhi