Third: Evaluate Data Quality Issues in the Data Provided

I evaluated data quality issues in the data provided using Python (fetch_assessment.ipynb), the following are the most prominent issues found in the data.

Receipts table:

- A significant number of columns in the receipts table have missing values, for instance bonusPointsEarned, bonusPointsEarnedReason, finishedDate, pointsAwardedDate, etc have around 50% of values missing.

- Missing values in key fields such as pointsEarned, purchasedItemCount and totalSpent makes it challenging to estimate brand popularity, segment customers and making data driven decisions.

Users table:

- There are more than 57% duplicate entries in the Users table, there seems to be an issue with capturing user data within the app.

- This redundancy could lead to inaccuracies in user data analysis and affect downstream processes when this table is joined with other tables. I would suggest removing the duplicate entries before loading data into the table.

Brands table:

- There are a lot of missing values (around 50%) for topBrand and categoryCode columns.

- category and categoryCode columns seem to have similar entries with categoryCode column having a lot more missing values, in fact four times than category. I would suggest that we standardize the categories and drop the categoryCode column.

- Recommendation: Instead of storing topBrand as a column within brands, would it be better store them in a separate look-up table for topBrands so that we can update the topBrands whenever necessary?