Fourth: Communicate with Stakeholders

Subject: Data quality issue report with Receipts, Brands, and Users data

Dear Sir/Madam (replaced by the product/business leader name),

After careful exploration on Receipts, Brands and Users datasets, I propose the following suggestions so that our company could have a more comprehensive and tidier dataset for future analysis to guide our business decision-making:

1. Discuss with data engineer team recording table design of Receipts: rewardsReceiptItemList
2. Collect more complete datasets because a significant portion of the data is missing
3. Delete duplicated records in Users table
4. Correct the anomalous relationship of barcode in Brands table and timestamp in Receipts table
5. Verify the correctness of abnormal values in Receipts table
6. Discuss with business team and redesign the category and category code in Brands table
7. Change the date format to mm/dd/yyyy

And my reasons are as follows:

1. **Inconsistent data structure**
   In current Receipts data, the Data format of rewardsReceiptItemList is not consistent, so it is not able to process in batch. For example, some items only have 6 features such as description, quantity, price, etc., while some other items have 18 features. This will lead to a lot of missing values when we want to structure the data and make effective analysis. We can discuss with data engineer team to see how to solve the issue.
2. **Missing data**
   All the three datasets include certain amounts of missing values. However, there are a significant number of missing values in Receipts and Brands tables. With the missing values over certain percentage, these features will not provide enough information. I have listed the attributes of the two tables with missing percentage here:

Table 1: Receipts Data

	Number of missing values	Missing value percentage (%)
pointsAwardedDate	582	52.01
bonusPointsEarnedReason	575	51.39
bonusPointsEarned	575	51.39
finishedDate	551	49.24
pointsEarned	510	45.58
purchasedItemCount	484	43.25

purchaseDate	448	40.04
rewardsReceiptItemList	440	39.32
totalSpent	435	38.87

Table 2: Brands Data

	Number of missing values	Missing value percentage (%)
categoryCode	650	55.70
topBrand	612	52.44
brandCode	234	20.05
category	155	13.28

We may have to collect more complete data to conduct a more detailed and insightful analysis

3. **Duplicated data**
   In the total 495 records of Users data, only 42.8% User_id are unique. Because my team need to merge data with other tables, the redundancy in Users data will lead to duplications in the merged table and thus we should avoid in advance.
4. **Anomalous relationships in Barcode and Date**
   a. The barcode in Brands table is used as identifier to merge data. However, I found that some products with same barcode represent different products. For example, barcode 511111605058 represents magazine brand and dairy brand at the same time. And it will lead to confusions when joining data.
   b. After transforming timestamp in Receipts table to date format, I found that some date anomalies. For example, some customers scanned receipt on 01/03/2021, while their purchase dates are 02/03/2021. We'd better check which stage led to this problem and try to avoid same problems.
5. **Abnormal values in Receipts table**
   According to some boxplots I drew, I found that some abnormal values in purchased item count and points earned.
   a. If customer didn't purchased item, there should be no receipt record, but our current Receipts data includes these invalid records
   b. For the points earned, some customers earned points over 5000 with one receipt. It's not very common in reality, especially when spending is less than $100. It's potential fraud in this case.
   c. There are potential fraud users in data. For example, one customer scanned 436 times, which is significantly higher than the other customers. In view of our business fairness, we can use some statistical methods to detect.
6. **Category & Category Code**
   Currently, there are some duplications in brand category. For example, "Dairy", " Dairy & Refrigerated" are same categories. In category code, the duplications are eliminated. However,

the category code misses some categories, such as ”Canned Goods & Soups” and “Deli”. I recommend that the business team design the category and category code again to solve the issues.

7. **Time Format**
   Current date format is UNIX Timestamp, which is hard to extract information. I suggest changing it to mm/dd/YYYY format or similar format so it can smooth date processing.

Please feel free to let me know if you have any questions. I’m happy to schedule a meeting to discuss these solutions in detail at your convenience.

Sincerely,

Shuwen Huang

(signature)