Fetch Rewards Coding Exercise - Data Analyst

1. **Review Existing Unstructured Data and Diagram a New Structured Relational Data Model**

Graphical user interface

Description automatically generated

**Model Structure and Data Preprocessing:**

The above diagram shows the structured relational data model I built for this exercise. By reviewing the unstructured JSON data, I found that the barcode needed for connecting receipts data schema with brands data schema is in the **rewardsReceiptItemList** column of receipts. Therefore, I used pandas JSON normalize function to expand items in **rewardsReceiptItemList** to multiple rows (with meta prefix set to ‘*receipt*\_’ and record prefix set to ‘*item*\_’) and built a new table **receipt\_items** with **(receipt\_id, item\_partnerItemId)** as the primary key. This table will extract all items from each receipt to answer the following questions.

1. **Write a query that directly answers a predetermined question from a business stakeholder**

The codes for the six questions are in the file *fetch\_reward\_cod.sql*.

* What are the top 5 brands by receipts scanned for most recent month?

Since the dataset only contains scanned dates in 2020/10 - 2020/11 and 2021/01 - 2021/03. The code for the most recent month is commented.

(Count all brands no matter the brandcode is in brands or not)

**HY-VEE**

**BEN AND JERRYS**

**PEPSI**

**KROGER**

**KLEENEX**

(Only keep brands with their brandcode in brands table)

**PEPSI**

**KLEENEX**

**KNORR**

**DORITOS**

**KRAFT**

* How does the ranking of the top 5 brands by receipts scanned for the recent month compared to the ranking for the previous month?

Skipped since dataset only contains scanned dates in 2020/10 - 2020/11 and 2021/01 - 2021/03.

* When considering average spend from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?
* When considering total number of items purchased from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

(Assume **rewardsReceiptStatus** Accepted is Finished)



* Which brand has the most spend among users who were created within the past 6 months?

HEMPLER’S

(Assume estimated spend is calculated by final price \* purchased item count. Since the current date is not defined, I counted all users and the code for past 6 months is in the comment)



* Which brand has the most transactions among users who were created within the past 6 months?

BEN AND JERRYS

(Assume one receipt is one transaction)



1. **Evaluate Data Quality Issues in the Data Provided**

* Duplicate data

There is duplicate data in users data schema where user id, created date, last login date, and other variables are completely the same. This could cause problems if a task requires to join the table because user id which should be the primary key cannot ensure uniqueness.

This can be resolved by deleting duplicated rows or setting an auto increment value to the user id.

* Missing data

The brands data schema is missing brand codes from the receipt items data. This does not have an impact on the data quality, but this may cause loss of information. From the first question above, the top 2 brands by receipts scanned are HY-VEE and BEN AND JERRYS, but neither of them is stored in the brands data schema. This could lead to loss of information when making decisions and strategies.

The missing data can be resolved by adding brand codes to brand data schema based on the brand codes in the receipt\_items.

* Ambiguous data

(Unsure) When I was trying to convert data frame of receipts data schema to csv in python pandas, the output csv has more rows than the data frame because some of the records (with many items) in **rewardsReceiptItemList** were separated to new rows. I tried to set *quoting* and *sep* in the function, but it still didn’t work. There could be potential format issues if the **rewardsReceiptItemList** contains ‘\n’ charactors.

When working on question 5 above, I found that the brand code is stored in the item list column of receipt and the total spend (per receipt) is stored in the receipts data schema. Therefore, one receipt has only one total spend value but has multiple items in the list and each item has a corresponding brand code. Currently I am using final price \* quantity purchased as an estimated value of the spend for each brand in a receipt. But it will lead to inaccuracy when calculating spend for each brand.

This can be resolved by adding another value in the **rewardsReceiptItemList**  indicating the total spend for each set of items in the same brand.

1. **Communicate with Stakeholders**

Dear Business Leaders and Stakeholders,

I hope this message finds you well.

This is Jiadong Li. I am working on your questions using receipts, users, and brands data schema, and I found several data quality issues that I believe need to be addressed to prevent losses and misinformation in the future.

First, when working on the top 5 brands by receipts scanned for most recent month, I found that some brands in the receipts are not stored in the brands data schema. This could cause a lack of information if we make decisions based on brands data in the future. I would suggest an update on the brands data based on the brands that appeared in the receipts.

Secondly, there are duplicated records in the users data schema with completely the same information. This can be resolved by deleting all duplicated records while keeping one of them; or adding any additional information in users data that can used by unique identifiers.

Another data quality issue I found is when calculating the spend by brands, the items on each receipt may come from different brands but there is no corresponding spend value for the items. I calculated an estimated spend but it is not accurate. I would suggest adding another value showing the spend for each item on the receipt.

Thank you for your time and if you have any questions, please kindly let me know.

Best,

Jiadong Li