**FETCH REWARDS CODING EXERCISE – DATA ANALYST**

**Interview Case Assignment Project**

**Group 1: Aakash Mundra, Aayush Jhingan, Yashika Chauhan**

**Course: BAN 5501 - DATA MANAGEMENT & SQL FOR ANALYTICS  
 Clark University  
 Instructor: Kathy Fogel  
 Date of Submission: 30th April 2023**

## 

Logo, company name

Description automatically generated

**First: Review Unstructured Data and Diagram a New Structured Relational Data Model**

The simplified relational diagram is the next slide has Four main tables:

1. **Transactions:** This table stores information about each receipt, Including the receipt ID, the date the receipt was issued, the bonus points earned for the purchase, and the user ID of the person who made the purchase.
2. **Brand:** This table stores information about each brand, including the brand ID and code.
3. **Items:** This table stores information about each OID, Bar code, descriptions of the product, the final price of each item, and their individual prices item price along with various information related to the item’s dataset. We witnessed lots of data in the items table that needed to be included.
4. **Users**: This table stores information about each user, including the user ID, state, last login information, etc. The Receipts table and the Brands table are connected via brand and bar codes, allowing us to track sales. The Users table and the Receipts table are connected via the user ID field, which allows us to track which user made each purchase.

ERD Diagram 


**Second: Write a query that directly answers a predetermined question from a business stakeholder**

**Question 1: What are the top 5 brands by receipts scanned for the most recent month**?

Ans:

Select "name", receipts\_scanned

From (

select b."name",

(select count(\*)

from "trans" as t, "item" as i

where b."brandcode" = i."brandcode"

and i."oid" = t."oid"

and t."datescanned" >= '2021-01-25 00:00:00-05') as receipts\_scanned

from "brands" as b

) as subquery

Where receipts\_scanned is not null

And receipts\_scanned > 0

Order by receipts\_scanned desc

Limit 5;

**Question 2: How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?**

Ans:

Select "name", receipts\_scanned

From (

select b."name",

(select count(\*)

from "trans" as t, "item" as i

where b."brandcode" = i."brandcode"

and i."oid" = t."oid"

and t."datescanned" >= '2020-12-25 00:00:00-05'

and t."datescanned" < '2021-01-25 00:00:00-05') as receipts\_scanned

from "brands" as b

) as subquery

Where receipts\_scanned is not null

And receipts\_scanned > 0

Order by receipts\_scanned desc

Limit 5;

**Question 3: When considering average spend from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?**

Ans:

Select AVG(totalSpent) From Receipts

Where rewardsReceiptStatus = “Accepted”;

*Output: Average spend on receipts with rewards status “Accepted”*

Select AVG(totalSpent) From Receipts

Where rewardsReceiptStatus = “Rejected”;

*Output: Average spend on receipts with rewards status “Rejected”. We compare the values returned by the above statements to find the greater among the two.*

## **Question 4: When considering total number of items purchased from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?**

Ans:

Select COUNT(purchasedItemCount) From Receipts

Where rewardsReceiptStatus = “Accepted”;

*Output: Count of items purchased on receipts with rewards status “Accepted”*

Select COUNT(purchasedItemCount) From Receipts

Where rewardsReceiptStatus = “Rejected”;

*Output: Count of items purchased on receipts with rewards status “Rejected”. We compare the values returned by the above statements to find the greater among the two.*

# **Question 5: Which brand has the most spend among users who were created within the past 6 months?**

Ans:

Select Users.user\_id, Users.createdDate, Receipts.receipt\_id as rid, Receipts.totalSpent, Transaction.receipt\_id as tid, Transaction.brand\_id as tbid, brand.\_id  
From((( Users

Join Receipts on Users.user\_id = Receipts.userId) Join Transactions on tid = Receipts.receipt\_id) Join Brands on Brand.brand\_id = tbid)

Top (totalSpent)  
Where createdDate = Dateadd(Month,datediff(Month,0,Dateadd(m,-6,current\_timestamp)),0)

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

Ans:

Select Users.users\_id, Users.createdDate, Receipts.receipt\_id as rid, Receipts.totalSpent, Transaction.receipt\_id as tid, Transaction.brand\_id as tbid, brand.brand\_id  
From((( Users  
Join Receipts on Users.user\_id = Receipts.userId) Join Transactions on tid = Receipts.receipt\_id) Join Brands on Brand.brand\_id = tbid)

Count (rewardsReceiptItemList)  
Where createdDate = Dateadd(Month,datediff(Month,0,Dateadd(m,-6,current\_timestamp)),0)

**Third: Evaluate Data Quality Issues in the Data Provided**

Ans:  
To review the data, we used the pandas library to read in the "trans.csv" file in Python and then checked the type of each feature in the table using the "dtypes" function.

Users: There are two data quality issues that have been identified:  
  
More than 50% of the total 495 user records are duplicates.

There is a minor problem with missing values in the signUpSource and state columns.

Brands:

The 'Baking' category owns most of the brands. No significant data quality problems have been detected, except for a significant amount of missing data in the 'topBrand' and 'categoryCode' columns.

When working with data, we made sure to check for any invalid or incomplete information. To do this, we used GROUP BY SQL operations to compare the values of certain fields to our expectations. Unfortunately, we discovered that there were many rows with inaccurate or missing data. In particular, the Receipt table had some items without a barcode or with an "ITEM NOT FOUND" description, and there were also redundancies in the various price columns. Additionally, there were some duplicate columns in the User table that caused problems with foreign key creation.

**Fourth: Communicate with Stakeholders**

Ans:

Subject: The data quality of Receipts, Users, and Brands data has problems.

Hi Jeff,

I conducted an exploratory analysis of the Receipts, Users, and Brands data and found some data quality issues that I believe are important to bring to your attention. Here's what I discovered:

Certain fields have a considerable amount of missing data, including finishedDate, pointsEarned, purchasedItemCount, totalSpent, rewardsReceiptItemList, topBrand, and categoryCode. These missing values may cause problems in deciding if users qualify for special offers or bonus points or in determining the points earned for transactions.

Some values in the pointsEarned, purchasedItemCount, and totalSpent columns appear to be out of place, and I recommend investigating the app's processes that produce these values to determine if they are legitimate or the result of errors.

More than half of the Users data has duplicate records, which should be eliminated to ensure there are no more anomalies in the future. The date formats are inconsistent with the standard MM/DD/YYYY or similar formats, and I suggest reviewing the database to ensure that the date fields are consistently captured and stored.

To tackle the data quality issues, we require a data cleansing process that guarantees completeness, accuracy, and consistency of the data. We also intend to enforce unique constraints for each dataset and establish foreign keys to establish the relationship between datasets, ensuring consistent and accurate data across all tables.

Furthermore, to maximize the value of our data, we require more information on customer demographics and purchase history. It is particularly essential to fill in the gaps and correct incomplete data. The additional information will enhance the precision and effectiveness of our analysis, enabling us to devise more targeted marketing campaigns and improve our product offerings.

I have a plan to address these data quality issues, and I'd like to discuss it with you further. Please let me know when a good time would be to schedule a meeting.

Regards,

Group 1 (Aakash Mundra, Aayush Jhingan, Yashika Chauhan