Fetch Rewards Coding Exercise - Analytics Engineer

Hardik Sandeep Fulfagar
fulfagarhardik@gmail.com

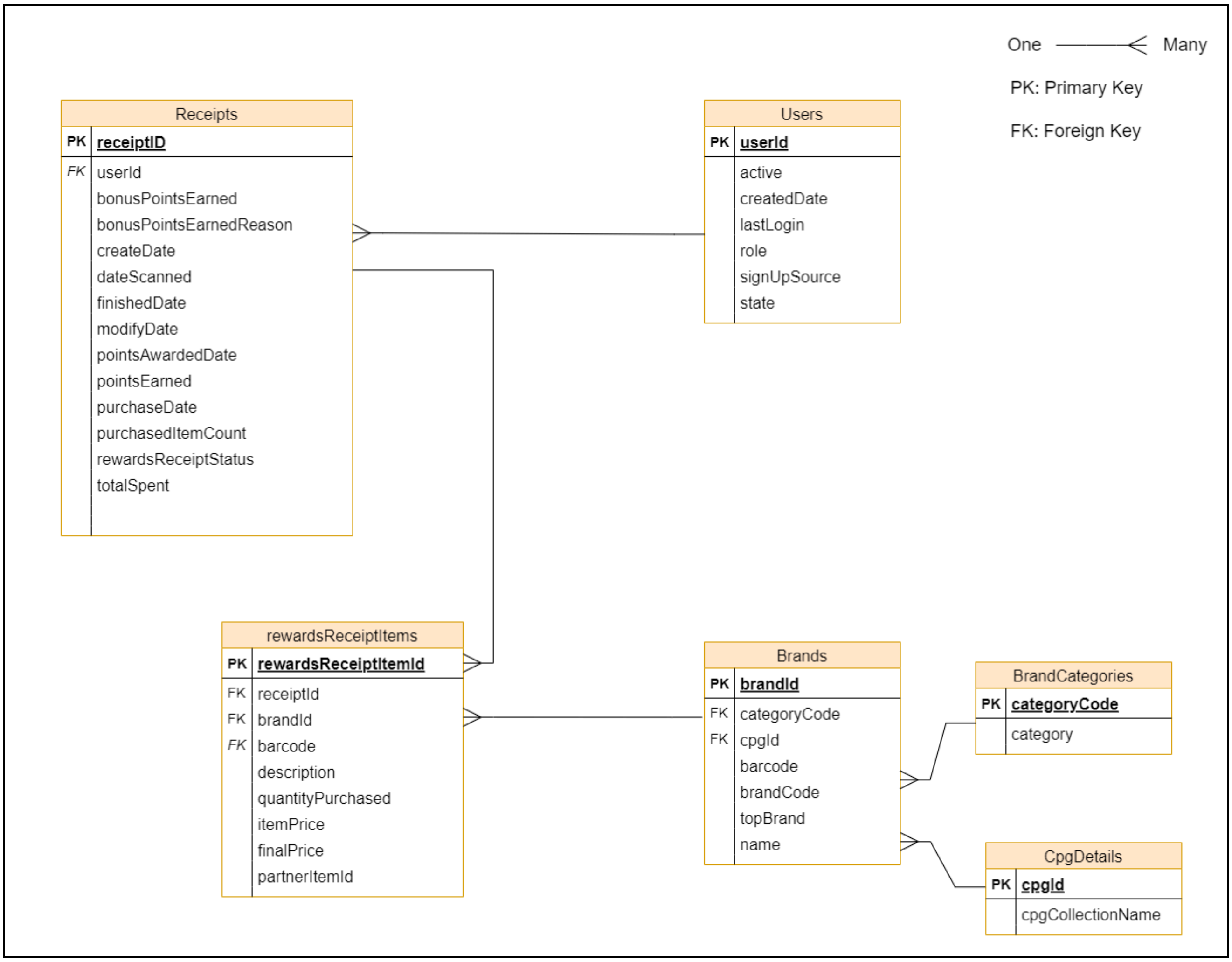
All the answers to each section of the assessment are provided in this PDF. All the supporting material, including codes and diagrams, is provided in the repository. Please refer to the repository if needed.

Table of Contents

First: ER Diagram

Task: To Review Existing Unstructured Data and Diagram a New Structured Relational Data Model

<u>ER Diagram of Proposed Structured Relational Data Model</u>



Second: Queries

Task: To Write queries that directly answer predetermined questions from a business stakeholder

Note: The SQL dialect that I have used for this coding exercise is ***SQLite****.*

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

```
WITH RecentMonth AS (
    SELECT
        MAX(purchaseDate) AS most_recent_date
    FROM
        Receipts
)
SELECT
    b.name AS brand_name,
    COUNT(r.receiptId) AS receipt_count
FROM
    Receipts r
JOIN
    rewardsReceiptItems ri ON r.receiptId = ri.receiptId
JOIN
    Brands b ON ri.brandId = b.brandId AND ri.barcode = b.barcode
JOIN
    CpgDetails c ON b.cpgId = c.cpgId
WHERE
    r.purchaseDate >= date('now', 'start of month', '-1 month') AND
    r.purchaseDate < date('now', 'start of month')
GROUP BY
    b.name
ORDER BY
    receipt_count DESC
LIMIT 5;
```

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

```
WITH AvgSpend AS (
    SELECT
        rewardsReceiptStatus,
        AVG(totalSpent) AS average_spend
    FROM
        Receipts
    WHERE
        rewardsReceiptStatus IN ('Accepted', 'Rejected')
    GROUP BY
        rewardsReceiptStatus
)


SELECT
    CASE
         WHEN (SELECT average_spend FROM AvgSpend WHERE rewardsReceiptStatus
= 'Accepted') >
              (SELECT average_spend FROM AvgSpend WHERE rewardsReceiptStatus
= 'Rejected')
        THEN 'Accepted'
        ELSE 'Rejected'
    END AS greater_average_spend;
```

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

```
WITH TotalItems AS (
    SELECT
        rewardsReceiptStatus,
        SUM(purchasedItemCount) AS total_items_purchased
    FROM
        Receipts
    WHERE
        rewardsReceiptStatus IN ('Accepted', 'Rejected')
    GROUP BY
        rewardsReceiptStatus
)
SELECT
    CASE
               WHEN (SELECT total_items_purchased FROM TotalItems WHERE
rewardsReceiptStatus = 'Accepted') >
                      (SELECT total_items_purchased FROM TotalItems WHERE
rewardsReceiptStatus = 'Rejected')
        THEN 'Accepted'
        ELSE 'Rejected'
    END AS greater_total_items_purchased;
```

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

```
WITH RecentUsers AS (
    SELECT
        userId
    FROM
        Users
    WHERE
        createdDate >= date('now', '-6 months')
)
SELECT
    b.name AS brand_name,
    SUM(ri.itemPrice * ri.quantityPurchased) AS total_spend
FROM
    Receipts r
JOIN
    rewardsReceiptItems ri ON r.receiptId = ri.receiptId
JOIN
    Brands b ON ri.brandId = b.brandId AND ri.barcode = b.barcode
JOIN
    CpgDetails c ON b.cpgId = c.cpgId
WHERE
    r.userId IN (SELECT userId FROM RecentUsers)
GROUP BY
    b.name
ORDER BY
    total_spend DESC
LIMIT 1;
```

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

```
WITH RecentUsers AS (
    SELECT
        userId
    FROM
        Users
    WHERE
        createdDate >= date('now', '-6 months')
)
SELECT
    b.name AS brand_name,
    COUNT(ri.rewardsReceiptItemId) AS transaction_count
FROM
    Receipts r
JOIN
    rewardsReceiptItems ri ON r.receiptId = ri.receiptId
JOIN
    Brands b ON ri.brandId = b.brandId AND ri.barcode = b.barcode
WHERE
    r.userId IN (SELECT userId FROM RecentUsers)
GROUP BY
    b.name
ORDER BY
    transaction_count DESC
LIMIT 1;
```

Third: Data Quality Issues

Task: To Evaluate Data Quality Issues in the Data Provided

While assessing the data quality issues, mainly three parameters are checked: *Completeness, Uniqueness, Consistency*.

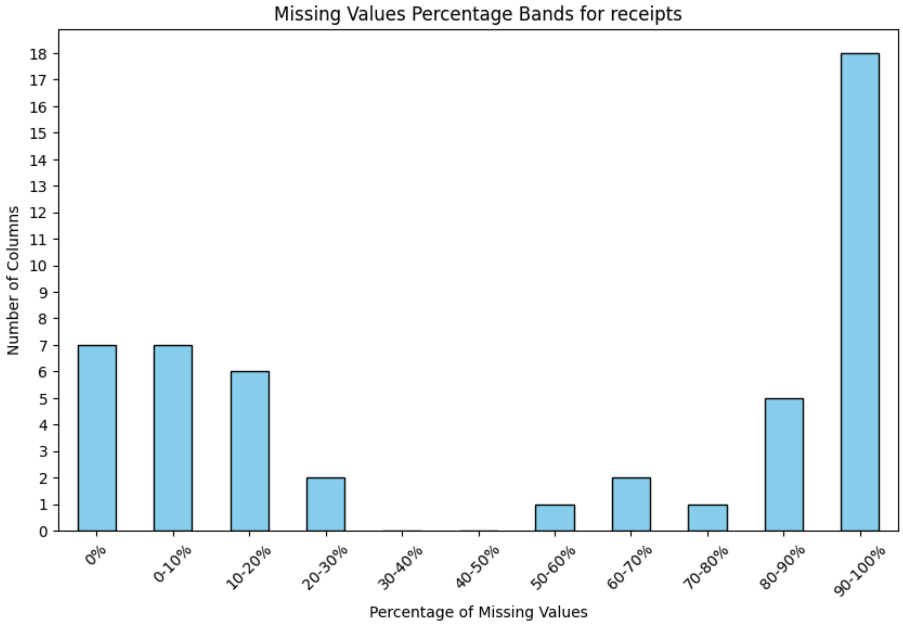
1. **Completeness**

   The inferences of checking for completeness of various data schemes are given below. Additionally, the missing values have been visualized in plots that categorize the columns based on the percentage of missing data they contain. These plots include bands (e.g., 0-10%, 10-20%) to show the frequency of columns within each band and a separate bar for columns with exactly 0% missing values, providing a clear overview of the missing data distribution.

   a. ***Receipts Data Scheme***

   This data scheme has significant missing data, especially in the rewardsReceiptItemList. Many fields in this section are populated for only a few records and are mostly absent for the majority of the records.

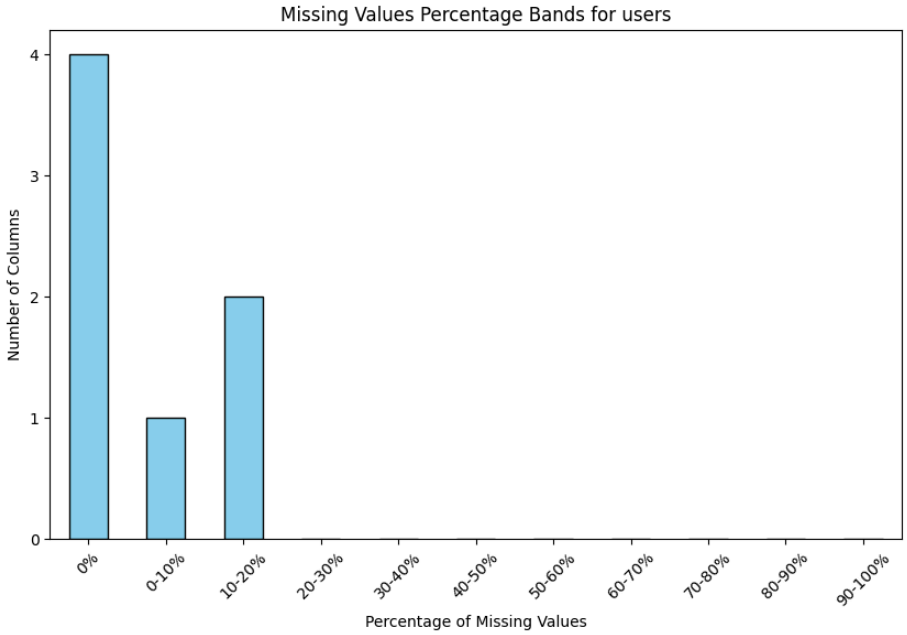
   <u>Plot for Receipts</u>



   b. ***Users Data Scheme***

   Most of the data in this scheme is complete, except for the lastLogin, signUpSource, and state columns. The signUpSource and state columns could be of particular business use.

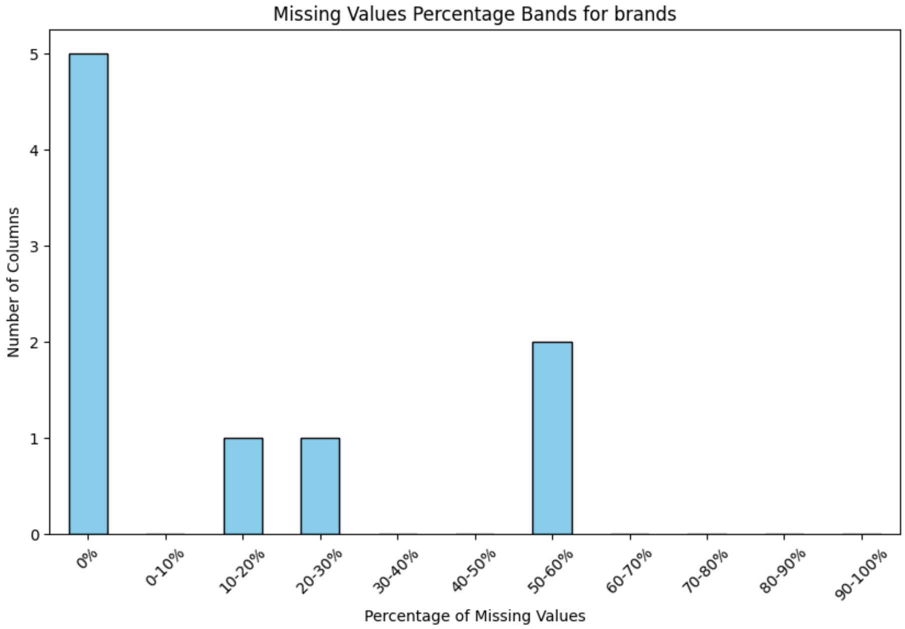
Plot for Users



c. ***Brands Data Scheme***

This data scheme has multiple columns with a significant number of missing values. Notable columns with missing data include brandCode and categoryCode.

Plot for Receipts



2. **Uniqueness**

   To check for uniqueness, the data schemes were examined for duplicate records. The findings are as follows:

   a. ***Receipts Data Scheme***
   There were 111 duplicate records found in the receipts data. Upon inspecting these duplicates, it was discovered that the data scheme is not consolidating items with the same barcode into a single record, despite having a separate quantity field. Instead, it is adding a new record for each instance of the item in the list.

   b. ***Users Data Scheme***
   This data scheme also has 283 duplicate records. This indicates that the same user has been recorded multiple times in the database, possibly due to the absence of a uniqueness check before adding a new user.

   c. ***Brands Data Scheme***
   No duplicate records were found in the brands data scheme.

3. **Consistency**
   After inspecting the given data, several consistency issues were identified:

   a. ***Brand Codes Discrepancy***
      - Number of brand codes in receipts not present in brands: 186
      - Unique brand codes in brands: 897
      - Unique brand codes in receipts: 227

   Inference
   There is a significant mismatch between the brand codes in the receipts and those listed in the brands data. Specifically, 186 brand codes appearing in the receipts are not present in the brands data, indicating potential discrepancies in brand code management or data integration issues.

   b. ***Barcode Discrepancy***
      - Barcodes in brands: 1160
      - Barcodes in receipts: 568
      - Number of barcodes in receipts not present in brands: 552

   Inference
   A substantial number of barcodes (552) found in the receipts are not present in the brands data. This discrepancy suggests possible issues with barcode standardization or incomplete data synchronization between receipts and brands datasets.

c. ***CPG ID Consistency***

There are 8 cpg_id values with more than one cpg_ref.

Inference

Multiple cpg_id values are associated with more than one cpg_ref, indicating inconsistent categorization or labeling within the data. This inconsistency can lead to confusion and errors in data analysis and reporting.

Fourth: Email

Task: Communicate with Stakeholders

Subject: Summary of Data Quality Evaluation and Proposed Next Steps

Hello,

I hope this message finds you well. I have completed an initial evaluation of our datasets and identified several quality issues, including missing values in key fields, duplicate records, and inconsistencies between datasets. Specifically, there are significant gaps in some fields in our receipts data, discrepancies in brand codes and barcodes between our receipts and brands data, and inconsistent *cpg_id* categorizations. Additionally, there are duplicate user records.

To address these issues effectively, I have a few questions:

1. What are the key business impacts of the missing data in critical fields? How should we prioritize filling these gaps?
2. Can you provide insights into current data entry processes to help understand the duplicates and inconsistencies in barcode and brand code data?
3. What guidelines should we follow to ensure consistency in *cpg_id* categorizations, and how important are these fields for our business analytics?

To resolve these issues, I need detailed documentation on the intended use and business rules for each data field, access to historical data to help fill in missing values and validate existing records, and feedback on proposed data validation rules.

Additionally, to optimize our data assets, I need clarification on how frequently data is updated and any seasonal patterns that might affect data quality. Insights into how the data is used across different departments would also help tailor our validation processes.

In production, I anticipate challenges with data volume and processing speed. To address these, we plan to optimize how we store and process the data, ensuring we can handle large datasets efficiently and maintain smooth operation.

Please let me know if you have any questions or need further details. I look forward to your feedback and am ready to discuss the next steps at your earliest convenience.

Best,
Hardik