**Fetch Rewards Coding Exercise – Response**

**1.Review unstructured JSON data and diagram a new structured relational data model**

**Issue**: All the 3 JSON files were showing EOF error

**Reason**: Improper manual editing or cut-off during file generation.

**Resolution:** Executed following python code to build clean JSON file

*\*\* 3 codes for each JSON FILE*

import json

with open("<file.json>", "r") as file:

for line in file:

try:

data = json.loads(line)

# process data

except json.JSONDecodeError:

print("Skipped malformed line:", line)

**Exported cleaned JSON file in csv and flattened the nested columns**

**Imported each csv file to DBeaver database, and created tables such as:**

**Receipts, Brands, UsersData**

\*\**Create table scripts are available in Github*

**Relationships**

* UsersData (1) ←→ (∞) Receipts
  + Relationship: One user can submit many receipts.
  + Join on: Users.userid = Receipts.userId
* Brands (1) ←→ (∞) Receipts
  + Relationship: One brand is linked to many receipts by barcode.
  + Join on: Brands.barcode = Receipts.barcode

**ER Diagram:**

A screenshot of a computer

AI-generated content may be incorrect.

**2.Generate a query that answers a predetermined business question**

**3.Generate a query to capture data quality issues against the new structured relational data model**

*The two questions are addressed collectively by standardising the data for each business query.*

NOTE: *Since there is no PurchaseDate, UserCreatedDate after 12/2021,hence recent date in the asked queries are taken as Max date. In Receipt table, valid barcodes are available only for month 2021-12, therefore previous month’s analysis was not possible*

***SQL file is available in GitHub link***

**4. Data Quality Observations**

**Question**: What is the authoritative source for each entity (e.g., receipts, users, brands)?

**Discovered data quality issue:**

* Logical mismatches, such as receipts linked to unknown or duplicate userid
* The rewardsReceiptStatus field appears to be stored with values like FINISHED instead of Accepted, and may need mapping for clarity.
* Inconsistent formats (e.g., malformed date objects, special characters)
* Some barcodes don’t match brand metadata, suggesting a need for further enrichment or validation. Valid barcodes would have helped to detect the outlier.

**What is Needed to Resolve These Issues:**

* Validation rules or schema definitions (e.g., required fields, expected ranges).
* Confirmation on which records should be considered authoritative or filtered out.
* Clarification on intended relationships across entities (e.g., one-to-many constraints).

**Information Needed for Optimization:**

* Business logic behind metrics (e.g., how “top brand” is defined — by total spend, count of receipts, or active users).
* Update frequency and data latency expectations.

**Performance & Scaling Considerations**

* Joins on high-cardinality fields (e.g., barcodes or user IDs) may impact performance.
* Need to design indexing strategies and possibly denormalize for dashboard/report speed.

**5.Write a short email or Slack message to the business stakeholder**

**Subject:** Key Insights from Receipt and Brand Data Analysis

Hi Team,

Following a detailed analysis of the latest receipts, user, and brand data, here are key findings to inform marketing strategy, product prioritization, and data quality assessment:

**1. Brand Engagement Trends**

* **Top 5 Brands This Month:** Based on receipt scans, brand **Tostitos** leads this month.
* **Month-over-Month Comparison:** Since valid barcodes with receipt are available only for one month, therefore previous month comparison is not available yet

**2. Spend Behavior by Reward Status**

* **Average Spend:** Receipts marked as **'Accepted'** (technically FINISHED) show **higher average spend** than those **'Rejected'**.
* **Total Items Purchased:** Similarly, **'Accepted'** receipts account for **more items purchased**, indicating more valuable transactions.

**3. New User Behavior (Last 6 Months)**

* **Highest Total Spend:** Brand **Tostitos** generated the most spend among newly created users.
* These brands may be performing well with newer customers, suggesting an opportunity for lifecycle marketing initiatives.

**Additional User analysis:**

User activity is heavily concentrated in Wisconsin, presenting an opportunity for the marketing team to drive adoption in other states through targeted and aggressive outreach. Additionally, most users sign up via email rather than through search engines, indicating potential to expand acquisition through SEO and paid search channels

Let me know if findings need to be embedded into a dashboard. Happy to walk through any of these in more detail.

Thanks