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

Created a star schema having Fact table as Rewards and other Dimension tables (Receipts, Brand,Users). Attached the image in repo

**Second: Write queries that directly answer predetermined questions from a business stakeholder**

**Used MySQL**

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

A computer code on a white background

Description automatically generated

* How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?
  + Similar to above code we will create 2 CTEs, one for recent month and another for previous month. Then we will compare the rank from these for top 5 brands
* When considering *average spend* from receipts with 'rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

A screenshot of a computer code

Description automatically generated

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

A screenshot of a computer code

Description automatically generated

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

A white screen with black text

Description automatically generated

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

A screenshot of a computer code

Description automatically generated

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

**Fourth: Communicate with Stakeholders**

* What questions do you have about the data?
  + The totalspend attribute does not have the value as sum of price for all the items in each receipt. Is it data error or there is some other calculation for totalspent?
  + How does receipt data corelate with brand data, is barcode used as join condition correct?
  + Receipts table have many null values, can those columns be deleted eg. originalFinalPrice, itemNumber
  + Having same values : preventTargetGapPoints, cpg\_ref. Are these required? Does the ref value in cpg has typo with values Cogs and Cpgs
  + What are the primary objectives or goals we aim to achieve with this data?
  + What specific insights or patterns are we hoping to uncover?
  + Are there key performance indicators (KPIs) or metrics that need to be tracked?
* How did you discover the data quality issues?

First formatted the file to remove whitespace and then checked for incorrect values ie error in the data itself. Next found missing data. Also, checked for range of values to see if there is any outlier. All the json files have record with attributes not in the same order.

* What do you need to know to resolve the data quality issues?
  + What specific data quality metrics or criteria are important to you?
  + Are there defined standards or expectations for data accuracy, completeness, and consistency?
  + What contextual information is required to understand and address the identified issues?
* What other information would you need to help you optimize the data assets you're trying to create?
  + What specific insights or analytics would be most valuable for your decision-making processes?
  + Are there additional data sources or contextual information that could enhance the value of the data?
  + How do you envision using the optimized data assets to drive business outcomes?
* What performance and scaling concerns do you anticipate in production and how do you plan to address them?
  + What are the anticipated volumes of data in production, and how frequently will it be updated?
  + Are there peak usage times or specific scenarios that may impact performance?
  + What measures are in place to ensure scalability and responsiveness in handling increased data loads?
  + How will the data infrastructure adapt to future growth or changes in data volume and complexity?