I reviewed the Json files attached and tried to convert it into a CSV file using python. I used json_normalize to get the data frame and convert it to CSV. While doing the above I came across a couple of problems:

1. The receipts list data is present inside the receipt data table and they are not relatable with receipts table.
2. Data types were changed and converted to string format.

To achieve the above I went ahead and used the Power query feature of Excel.

In this case we can directly import json file.One of the roadblockers for this method was the delimiter were not present in the Json file after each record and hence I had to manually edit the json file and add "," after the end of the data present in the file.

After the above I went ahead and created an Entity relationship diagram in MYSQL workbench.

Explanation of the ER diagram:

1. The Receipt table has 14 columns and ID is the primary key and user_id is the foreign key. The foreign key will link the users data table with the Receipts table.
2. User table has a total of 7 columns where ID is the primary key.
3. Brands has 8 columns where ID is the primary key.
4. Receipts list has 12 columns

SQL Queries:

1. What are the top 5 brands by receipts scanned for the most recent month?
    a. select name, count(*) from Receipts_list p join Brands b on p.Brands_id = b._id join Receipts_Data r on r._id = p.receipts_id where date_trunc('month', r.purchasedate) = max(date_trunc('month', r.purchaseDate)) order by count(*) desc limit 5
2. When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?
    a. Select avg(totalSpent) from Receipts_Data where rewardsReceiptStatus = "Accepted"
    b. Select avg(totalSpent) from Receipts_Data where rewardsReceiptStatus = "Rejected"
        i. We compare the above two queries to find the greater among them.
3. When considering *total number of items purchased* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?
    a. Select Count(purchaseditemCount) from Receipts_Data WHERE rewardsReceiptStatus = "Accepted";
    b. Select Count(purchaseditemCount) from Receipts_Data WHERE rewardsReceiptStatus = "Rejected";

        i. We compare the above two queries to find the greater among them.

4. Which brand has the most *spend* among users who were created within the past 6 months?
    a. Select Users._id,users.createdDate,Receipts_Data.receipts_id as tid, Receipts_Data.totalSpent, Receipts_List.rerceipts_id as rlid, Receipts_List.Brands_id as rlbid, Brands._id FROM (((users Join Receipts_Data on Users._id = Receipts_Data.user_id) join receipts_list on rlid = receipts_Data.receipt_id) Join Brands on Brands.brand_id = rlbid) Top(totalSpent) Where createdDate = Dateadd(Month,datediff(Month,0,Dateadd(m,-6,current_timestamp)),0)
5. Which brand has the most *transactions* among users who were created within the past 6 months?
    a. Select Users._id,users.createdDate,Receipts_Data.receipts_id as tid, Receipts_Data.totalSpent, Receipts_List.rerceipts_id as rlid, Receipts_List.Brands_id as rlbid, Brands._id FROM (((users Join Receipts_Data on Users._id = Receipts_Data.user_id) join receipts_list on rlid = receipts_Data.receipt_id) Join Brands on Brands.brand_id = rlbid) count(purchasedItemcount) Where createdDate = Dateadd(Month,datediff(Month,0,Dateadd(m,-6,current_timestamp)),0)

There are few steps that need to be followed to clean the data and point out issues in the dataset. Some of them are:

1. Check for duplicate data.
2. Check for data types.
3. Understanding the relationship of variables
4. Count the missing values

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

Hi,

I hope you are doing well. We're very thrilled to talk with you about the latest project developments. We appreciate your fear about the outcome; however, there is some information that we can give.

First and foremost, we needed to know what the end aim or business issue is with this data. We also need to identify how we'll measure the ultimate output, or key performance indicators.

Second, many entries in the Receipts tables were duplicated as I was going over the data. A few fields in the receipts data, such as bonus points gained, are not completed, and there are many null values. These columns have more than 40% missing values: item list description, fetch rewards review, and user flagged bar codes.

I was having trouble determining a few business challenges because of these missing values. In the brands table, the barcode data is not in the correct, category code needs accurate information for classifying the product data Users must be categorised, therefore additional information about them is required.

Furthermore, we require more information regarding brand code, as that column is deficient in many areas. While researching top brands, the bar code plays an important part. We've

almost completed the maximum number of predefined questions; the remaining portion will be selected after data preparation and quality assurance.
We appreciate that you are looking forward to the final presentation, but we wanted to double-check for data quality issues in order to achieve better outcomes. We are currently working on the final presentation. If you wish to meet before the production is finished, please plan a meeting. I will send out information about the modifications and final report schedule as soon as possible.

Thank you for your patience.

Respectfully,

Team analytics