**Fetch Rewards Assessment Documentation**

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

Data Given – Receipts (Json format) --- semi structured data, Users (Json format) --- semi structured data, Brands (Json format).

Firstly, I reviewed Json files and tried to convert using python. So, I used a python program which inputs in json file and flatten that file using (json\_normalize) to get the data frame and converted that to CSV.

Problems faced using this technique:

1. Data types were changed and converted to string format.
2. The receipts-list data that is present inside the receipts data table is not relatable with receipts table.
3. Using that csv file there is no way I can create a relational data base in MYSQL server.

One more Technique using power query (Excel):

Here we can directly import json file from data importing technique. But there is a problem here that is data in Json files doesn’t have proper delimiter (comma) that Excel can interpret.

I used Notepad++ to include comma after each data entry.

Loaded that data in to excel by using power Query I decided to open the data frame columns and convert it into tables.

I have done the same thing for all three tables and found out that there are four tables inside the database.

Graphical user interface, application, table, Excel

Description automatically generated

BRANDS DATA TABLE.

Graphical user interface, application

Description automatically generated

RECEIPTS DATA TABLE

Graphical user interface, application, table, Excel

Description automatically generated

RECEIPTS-LIST DATA TABLE

Graphical user interface, application, table

Description automatically generated

USERS DATA TABLE.

After this I have created a relational model to this data using MYSQL work bench.

Diagram

Description automatically generated

Explanation of this relational model:

1. Receipts has total 14 columns where ID will be the primary key. User-id the foreign key that will link to the user’s data table.
2. User table has total of 7 columns where ID is the primary key.
3. Receipts-List has 14 columns.
4. Brands has 8 columns.
5. CPG is derived from brands table and it has 2 columns.
6. **Generate a query that answers a predetermined business question.**

I have submitted a SQL file which contains all the queries to get the requirements.

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

I used python to load the existing csv files (Relational model) and understand the relation between the data.

There are few steps that we need to follow to understand the data quality and to point out the issues.

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

So, I have loaded all the converted csv files into python and did the following.

1. Count the null values and replace the null values with mean (for continuous variables), replace with mode (for categorial variables).
2. Removed the duplicates values.
3. Checked the data types of all three tables and replaced with strings, int, float, and date.
4. Date column is in UNIX format and needed to be replaced.
5. Bad Columns names.
6. More than 40% of null values in user’s data tables.
7. There are few data entry issues in brands table like few entries of bar code are wrong.

All the things that have done in python are in the google collab file (Data-quality-Issues).

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

Hello “Respected Stake Holder”,

I hope this email find you well. we are very excited to discuss with you regarding the updates on this project. We understand that you are concerned regarding the result, however we have certain information that can be provided.

Firstly, we wanted to know what is the final goal or business issue that we are dealing with this data? Along with that we also, we need to know how we are going to measure the final result, the key performance indicators.

Secondly, when I was going through data, in Receipts tables there were many entries that were repeated. There are few columns in receipts data such as bonus points earned reason column is not filled there are lot of null values. Item list description, fetch rewards review, user flagged bar codes- These columns have more than 40% of missing values. Due to these missing values we were facing issues to determine few business problems. We are just wondering if there is any other data source. If there is any, it would help us to get through data quality issues.

In brands table, brands bar code is not in correct format, category code needs accurate information for classifying the product data, Wrong entries in brand code. Users are needed to be classified so need more information regarding users.

Furthermore, we need more information regarding brand code, Because, there is lot of information missing in that column. While we are finding out regarding top brands bar code undergoes a key role. We almost solved maximum amount of predetermined questions, we need to undergo data preprocessing and quality check to determine the remaining part.

We understand you are very excited for the final presentation, but we wanted to make sure about data quality issues for better results. We are in the process of producing final presentation. Please schedule a meeting if you want to meet before finalizing the production. I will send out an information soon regarding the updates and final report scheduling.

Thank you for your patience.

Respectively,

Team analytics.