FETCH DATA ANALYTICS TAKE HOME ASSESSMENT

Data Modeling:

There are 4 .csv files which can be considered as 4 different entities of the Fetch Rewards System:

1) users.csv:

This contains the demographic details of the users who have been using the Fetch application for scanning and earning rewards. It contains their details like gender, birthdate, date when they created their Fetch profile, state where they live, the sign up source and platform, last time when they logged into the app and finally a unique identifier for each user.

2) brands.csv:

This contains the details of the brands associated with Fetch. It contains details like brand card and brand code, the CPG identifier, category of brand and its category code, name of the brand, its description, other brands to which it may be related and finally a unique identifier for each brand.

3) receipts.csv:

This contains the details of all the receipts scanned by users and uploaded on the Fetch app. It contains details like a unique identifier for each receipt scanned and uploaded, the store name where receipt was printed, purchase date, time, the date when the bill was scanned on Fetch, total money spent, status of rewards received, the user who purchased, quantity of items purchased, pending date, modify date, flagged date, processed date, finished date, rejected date, whether the receipt need further review from Fetch, if it is a physical receipt or digital receipt, deleted or not and non-point earning receipt status.

4) receipt_items.csv:

This is the biggest file and contains the details of all the items present in each receipt scanned by users and uploaded on the Fetch app and corresponding receipt ID. It contains details like a unique identifier for each receipt scanned and uploaded, the identifier for each receipt item, the item index, description of item, barcode and item brand code, quantity of each item purchased, the total money spent on each item, points earned by the user per item, rewards group, the original item name and modified date.

Based on the analysis, I have figured out:

A) Different keys in the model:

- There will be 4 entities in the E-R model - users, brands, receipts, and receipt_items.
- The ID column in users will be its primary key to uniquely identify each user.
- The ID column in receipts will be its primary key to uniquely identify each receipt and it will also act as the foreign key to the REWARDS_RECEIPT_ID column in receipt_items entity. The USER_ID column in receipts will act as a foreign key to refer to the ID in users table and refer it.

- The REWARDS_RECEIPT_ITEM_ID will act as the primary key of the receipt_items table because we can identify each row uniquely. The REWARDS_RECEIPT_ID will act as the foreign key to the receipts table.
- The ID column in brands will be its primary key to uniquely identify each brand.

Assumption/Ambiguous:

- The BARCODE column in the brands signifies the barcode of the brand. But the BARCODE column in the receipt_items may be ambiguous as it can be either the brand barcode or it can be the item barcode. Upon manual inspection, some values overlap in the columns and some do not. Also, there is a lot of difference in the lengths of barcodes in the receipt_items entity. So, we cannot clearly consider a connection.
- Again, the BRAND_CODE column in brands represents the brands and the BRAND_CODE in the receipt_items may be the brand code of the item purchased. But upon searching a few entries of brand codes present in the receipt_items, they were not visible in the brands table. So, a foreign key relationship cannot exist.

B) Cardinality:

- A one-to-many relationship exists between the users and the receipts table because 1 user can have 0 or more receipts.
- A one-to-many relationship exists between the receipts and the receipt_items table because 1 receipt can have 0 or more items.
- A one-to-one relationship exists between the brands and receipt_items because 1 item in the receipt_items can belong to just 1 specific brand.

Business Queries Solution:

I have used the PostgresSQL database and SQL language for creating the database, tables and running the queries.

1) Which brand saw the most dollars spent in the month of June?

Ans)

This query was ambiguous and as explained before due to unclear common columns between brands and receipt_items column, it will be difficult to get an answer.

2) Which user spent the most money in the month of August?

Ans)

For this, since we need to find the expenditure of the user we can use the receipts table alone as it has details of the user and their spends.
Firstly, I have filtered the data based on the PURCHASE_DATE. I chose this column because we have been asked about the money spent in August. This will be equivalent to the money spent when

the user purchased items. I have filtered based on PURCHASE_DATE where the month is 8 (August). On this, I have calculated the sum of total money spent by the user using the TOTAL_SPENT column and grouped it by each user based on USER_ID column. Finally, I have returned 1 row as we want the top expenditure.

3) What user bought the most expensive item?

Ans)

For this, since we need to find the user who purchased the most expensive item which can again be found from the receipt_items table.
I have used the concept of subquery and join. Firstly, we need to find the maximum value of money spent on the item which is possible using the TOTAL_FINAL_PRICE column and the QUANTITY_PURCHASED column. We need to divide the TOTAL_FINAL_PRICE by QUANTITY_PURCHASED to get the price of each unit of the item. We can find the maximum value using the above calculation from all the items purchased for a single user. I have retrieved the top 1 REWARDS_RECEIPT_ITEM_ID of the user which satisfies the above calculation. Using this, REWARDS_RECEIPT_ITEM_ID in the outer query. To get the user id, we need to join the receipts and receipt_items table as receipts has the USER_ID column. From the retrieved REWARDS_RECEIPT_ITEM_ID, we can fetch the user from the receipts after combining the tables as REWARDS_RECEIPT_ITEM_ID is unique.

4) What is the name of the most expensive item purchased?

Ans)

For this, since we need to find the name of the most expensive item which can again be found from the receipt_items table.
I have used the concept of subquery. Firstly, we need to find the maximum value of money spent on the item which is possible using the TOTAL_FINAL_PRICE column and the QUANTITY_PURCHASED column. We need to divide the TOTAL_FINAL_PRICE by QUANTITY_PURCHASED to get the price of each unit of the item. We can find the maximum value using the above calculation from all the items purchased for a single user. I have retrieved the top 1 REWARDS_RECEIPT_ITEM_ID of the user which satisfies the above calculation. Using this, REWARDS_RECEIPT_ITEM_ID in the outer query. I have fetched the distinct item using the ORIGINAL_RECEIPT_ITEM_TEXT column.

5) How many users scanned in each month?

Ans)

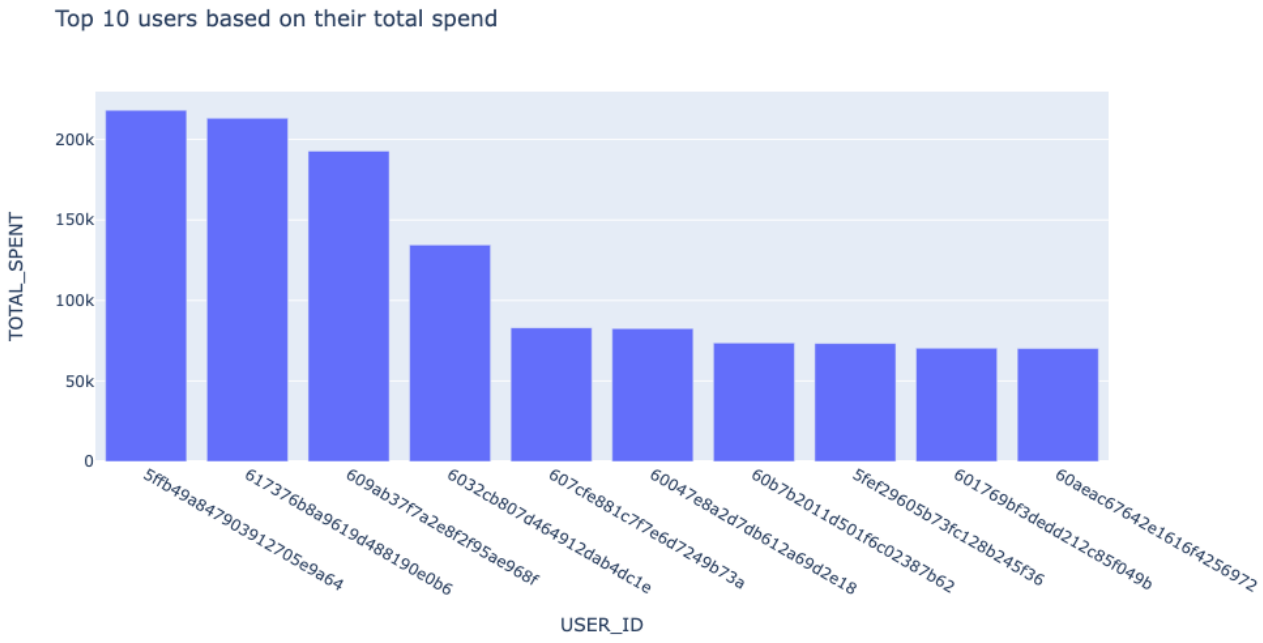
For this, since we need to find the users scanned each month, we need to count the number of users who scanned bills for each month. This can be ambiguous and can have 2 solutions - one with just the number of users who scanned and second with unique users who scanned each month. The second approach is better as it helps us understand clearly how many different users have scanned bills.
Firstly, I have grouped the data based on the DATE_SCANNED from receipts table. I chose this column because we have been asked about the user scans. I have extracted the month from the

DATE_SCANNED and counted the number of distinct users who scanned. Finally, I have returned the counts grouped by months.

Something noteworthy about the data and its interpretation:

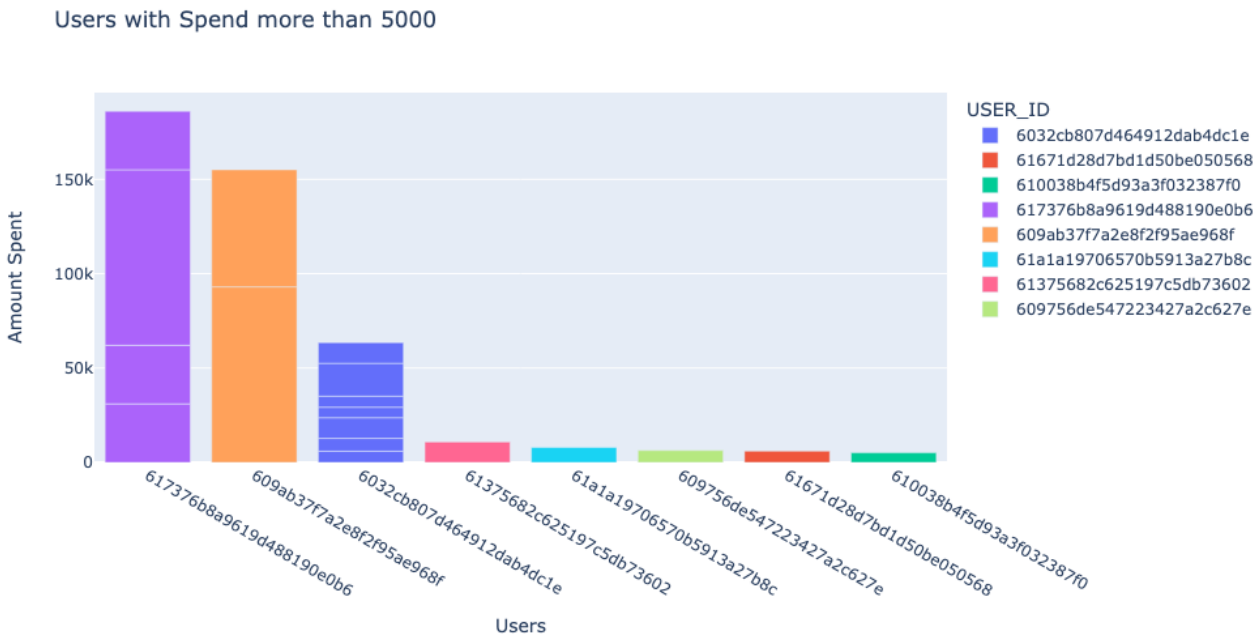
1) Top 10 Users based on Total Spent

It is important to understand, who are the top users who are spending money on buying items and scanning receipts on a frequent basis. **This can help the management to decide a way to devise promotions or extra rewards for the users who spend more which indirectly earns more revenue.**
I have used the TOTAL_SPENT feature in the receipts entity to calculate the total money spent and grouped the sum per user in a descending order. Of the rows available, I have filtered the first 10 rows which represents the top 10 users based on the Total Spent feature.



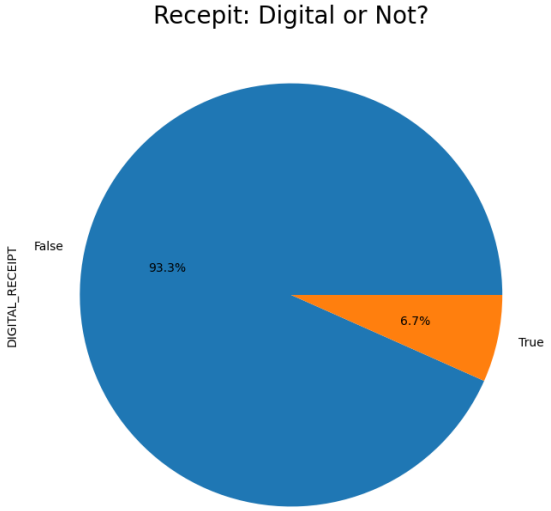
2) Users with Total Spend more than 5000

It is important to understand, who are the users who are spending more than 5000 on buying items. **This can help the management to understand which customers are spending more and they can be given more offers in terms of flashbacks or discounts.**
I have filtered the dataset based on the TOTAL_SPENT feature with condition that its total sum for the user must be more than 5000. This will count the sum of money spent based on all the transactions done by the user irrespective of date of purchase. We can clarify it more by adding the date filter.



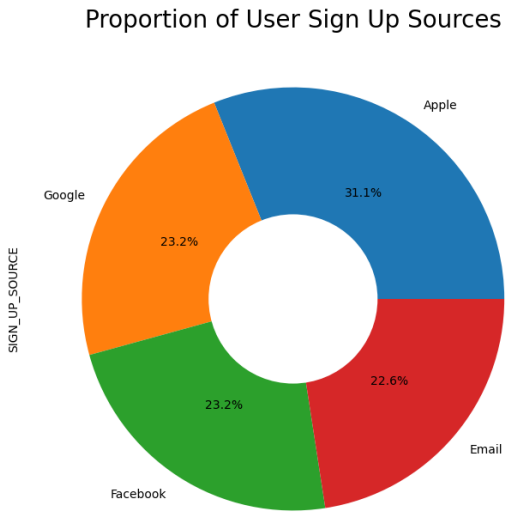
3) Digital Receipts or not?

This visualization can help the stakeholders understand whether the users/people prefer scanning paper receipts or e-receipts. It can be seen from the chart that the close to 93% of the receipts scanned are Non-Digital - paper receipts. **This means that people prefer shopping and scanning them rather than e-receipts. It also highlights that people are not willing to link their Amazon or PayPal accounts to the Fetch App which allows to scan the e-receipts.**
I have filtered the dataset based on the DIGITAL_RECEIPT feature and counted the number of values in the True and False category. The False represents Non-Digital receipt.



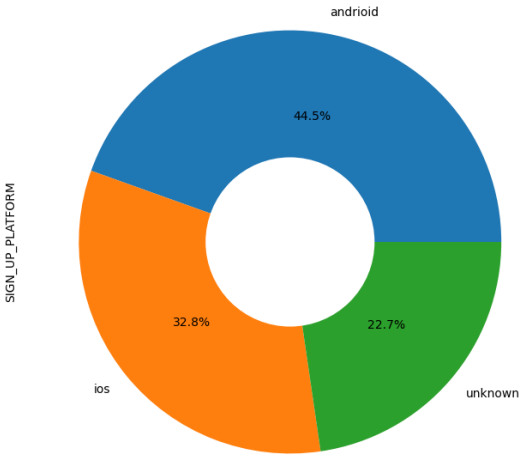
4) Proportion of User Sign Up Sources:

This visualization can help the stakeholders understand what are the different sources and the proportion of people who have signed up to the Fetch app using which source. It can be seen that Apple is leading as a sign up source with Google, Email and Facebook contributing approximately equally. **This means that people are more reliant on Apple source. Also, management can estimate whether to introduce more sources for sign up like Instagram, Snapchat etc.**
I have filtered the dataset based on the SIGN_UP_SOURCE feature from users and counted the number of values in the available categories.



5) Proportion of User Sign Up Platforms:

This visualization can help the stakeholders understand what are the different platforms prevalent among the masses and the proportion of people who have signed up to the Fetch app using which platform. It can be seen that Android is the top choice of people (approximately 45%) with iOS platform nearly 33%. There is a third category of ‘unknown’ which may be different from Android and iOS or it may be blank and replaced in the table with ‘unknown’ field. **This means that people are more reliant on Android platform which is slightly contradicting the previous visualisation which says Apple is the common sign up source.**
I have filtered the dataset based on the SIGN_UP_PLATFORM feature from users and counted the number of values in the available categories.


Proportion of User Sign Up Platforms
andrioid
44.5%
SIGN_UP_PLATFORM
32.8%
22.7%
ios
unknown