2. Queries that directly answer predetermined questions from a business stakeholder

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

Answer: Removing the "ITEM NOT FOUND" entries, the top 5 brands scanned for the month of Feb-2021 (the most recent month) are

1. FLIPBELT LEVEL TERRAIN WAIST POUCH, NEON YELLOW, LARGE/32-35

2. THINDUST SUMMER FACE MASK - SUN PROTECTION NECK GAITER FOR OUTDOORACTIVITIES

3. MUELLER AUSTRIA HYPERGRIND PRECISION ELECTRIC SPICE/COFFEE GRINDER MILLWITH LARGE GRINDING CAPACITY AND HD MOTOR ALSO FOR SPICES, HERBS, NUTS,GRAINS, WHITE

4. DELIMEX

5. GOOD SEASONS

- How does the ranking of the top 5 brands by receipts scanned for the recent month compare to the ranking for the previous month?

Answer: The top 5 brands scanned for the month of Jan-2021 are

1. *DIGIORNO CHEESE*

2. COOL WHIP

3. GODIVA INSTANT PUDDING MIX

4. DEVOUR

5. F. WHITLOCK & SONS BBQ SAUCE

The more popular brands scanned in Jan-2021 lie in the Food and Dining brands, whereas Feb-2021 brands are mixed with retail and dining

- When considering *average spend* from receipts with 'rewardsReceiptStatus' of 'Accepted' or 'Rejected', which is greater?

Answer: For rewardsReceiptStatus = 'FINISHED' (meaning ACCEPTED), the average spends (80.9) is higher than the average spends of rewardsReceiptStatus = 'REJECTED' (23.3)

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

Answer: For rewardsReceiptStatus = 'FINISHED' (meaning ACCEPTED), the total number of items purchased (8184) is higher than total number of items purchased of rewardsReceiptStatus = 'REJECTED' (173)

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

Answer: Brand with most spends: PULL-UPS (4647)

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

Answer: Brand with most transactions: DIGIORNO CHEESE (642)

3. Evaluation of Data Quality Issues in the Data Provided

- **Non-uniqueness of ids in users table**: of The number of unique users in the users table is 212 only. However the users table has 495 records. Ideally the table must be unique in terms of id column. Upon further inspection it was found out that theses are duplicated rows (other column values being the same). A unique key contraint would have helped to avoid that.
- **Null values in users table:** The fields of *lastLogin, signUpSource,* and *state* have null value percentage of 12.5%, 9.7% and 11.3% respectively.
- There are 117 users who have entries in the receipts table but are not present in the users table. Ideally, the userIds in receipts table should be a subset of all user ids present in the users table.
- The nested JSON structure for *rewardsReceiptItemList* field is not consistent. The keys change as per the value of the *rewardsReceiptStatus* field. For creating a relational database, we should have keys which are consistently present (even if the corresponding values are null values), and should be collectively exhaustive of all scenarios (example, when the status of the receipt is SUBMITTED, FLAGGED, or REJECTED).
- There should be a separate data source which contains the unnested keys of *rewardsReceiptItemList* field. This way, we can get analytical insights on a brand level in a better way. The volume of this table would be very high so we should think of creating indexes and partitions which would optimize analytical queries for brand level analyses.

- Just as *rewardsReceiptStatus* is present on a receipt level, the status should also be present on each item level so that we are able to identify the individual status of each item in the receipt.
- There are 3 cpg_ids which are present in the receipts table (rewardsProductPartnerId) but not present in the brands table. Ideally this should be a subset of the cpg_ids of brands table. This is a process flow that should be set up as part of brand onboarding