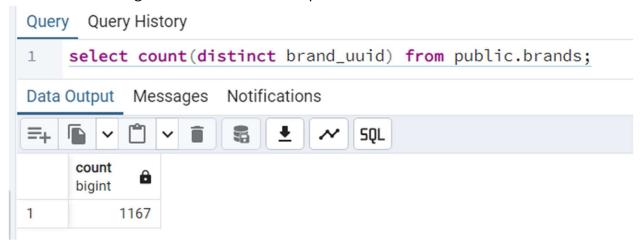
DATA QUALITY REPORT

Methodology

1. Checking the quality of the data in individual tables.

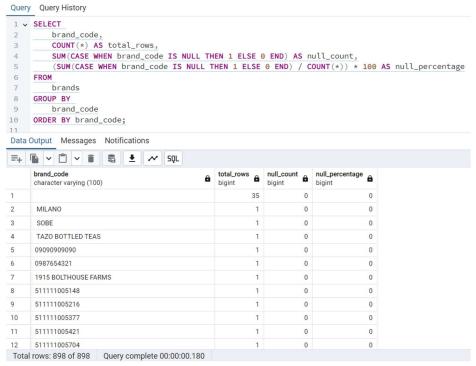
Brands Table

a. Checking the uniqueness of the brand_uuid as it's supposed to be a unique identifier according to the limited metadata provided:



Result: Brand_uuid is a unique column.

b. Checking for nulls in each column based on the importance of the column:



Result: Brand_code which is an extremely important foreign key for the brands table connecting it to the item receipts table has nulls, blank values, inconsistencies (text codes as well as numeric codes).

Fix: We should come up with a nomenclature for brand codes that should follow a consistent format with no missing values or nulls.

c. Data redundancy: I found quite a few columns with repetitive information, hence we could eliminate those columns to avoid redundancy.

Users Table

a. Checking for uniqueness of the unique identifier, user_id:

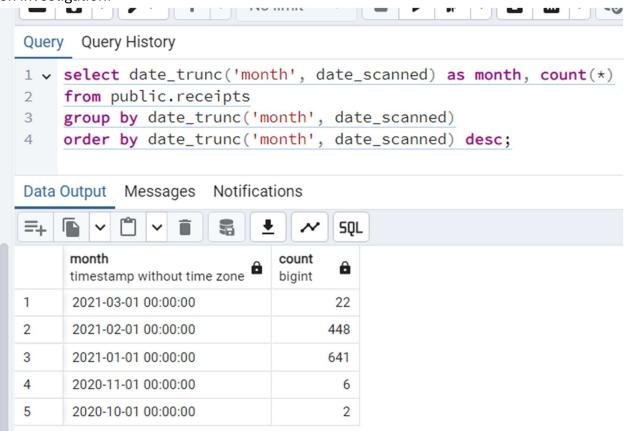
A similar query to the one above found that there are duplicate values in this column and in turn duplicate rows in the users table with only 212 unique values out of the 500+ records in the users table.

Receipts Table

- a. The bonus_points_earned_reason column can be streamlined to be broken up into further columns or made to follow a discrete value selection column to support further analysis.
- b. There are redundant columns with date_scanned and finished_date having the

same values.

c. The distribution of receipts scanned amongst different months is abornmal based on investigation:



Result: Either the data is not being completely captured or there could be business reasons for this.

d. Interesting find:

There are 258 unique users in the receipts table but only 212 unique users in our users table. Hence we are not capturing all the information in our users table or

there is something wrong in the receipts table:

Query Query History

- 1 v SELECT user_id, count(*) FROM public.receipts
- group by user_id
- 3 order by count DESC

Data Output Messages Notifications

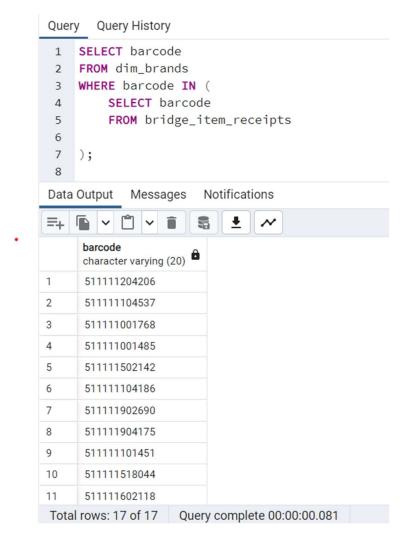
=+		<u>*</u> ~ 5
	user_id character varying (24)	count bigint
1	5fc961c3b8cfca11a077dd33	436
2	59c124bae4b0299e55b0f330	58
3	54943462e4b07e684157a5	50
4	5fa41775898c7a11a6bcef3e	21
5	5ff5d15aeb7c7d12096d91a2	20
6	600fb1ac73c60b12049027bb	16
7	5ff1e194b6a9d73a3a9f1052	14
8	5ff47392c3d63511e2a47881	10
9	600987d77d983a11f63cfa92	10
10	5a43c08fe4b014fd6b6a0612	9
11	6007464b6e64691717e8c1f0	8
12	5ff1e1eacfcf6c399c274ae6	8
13	5fff4beedf9ace121f0c17ea	8
14	5ff370c562fde912123a5e0e	7
15	6010bddaa4b74c120bd19dfb	7
16	5fff2698h3348h03eh45hh10	7

Item Receipts Table

a. The item receipts table has 31 columns which are highly dimensional in nature. To solve the business questions, I reduced the dimensionality by reducing the number of columns to only necessary 12 columns in the bridge_item_receipts table.

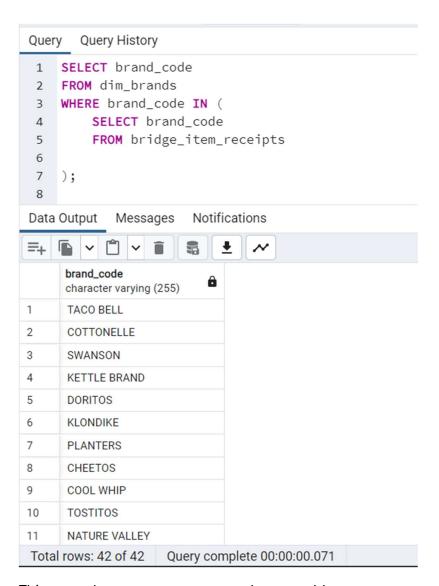
It did not have any primary key, so I created the item_receipt_id key in the table.

To obtain the foreign key relation between the bridge_item_receipts table and the dim_brands table, I checked the following which did not give sufficient outcome.

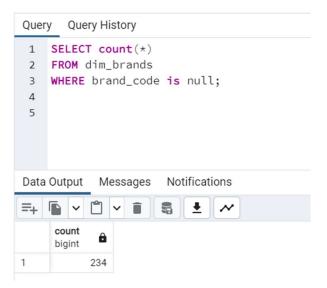


Barcode could not be created a foreign key.

Further, I checked the relation using brand_code as the key:

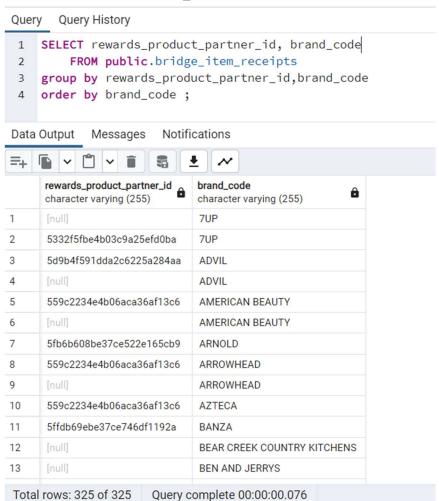


This was a better way to connect the two tables.



However, it had nulls and therefore to make it as Primary and Foreign Key, the nulls have to be removed or taken care of.

b. Insufficient information about the partner_item_id and rewards_product_partner_id and its relation with brand_code.



Query **Query History** SELECT rewards_product_partner_id, brand_code,partner_item_id 1 FROM public.bridge_item_receipts 2 group by rewards_product_partner_id,brand_code,partner_item_id order by brand_code ; Data Output Messages Notifications =+ rewards_product_partner_id brand_code partner_item_id character varying (255) character varying (255) character varying (255) 1 5332f5fbe4b03c9a25efd0ba 7UP 1432 2 7UP 1422 3 5332f5fbe4b03c9a25efd0ba 7UP 1456 4 5d9b4f591dda2c6225a284aa **ADVIL** 1188 5 5d9b4f591dda2c6225a284aa **ADVIL** 1190 6 5d9b4f591dda2c6225a284aa **ADVIL** 1187 7 5d9b4f591dda2c6225a284aa **ADVIL** 1180 8 5d9b4f591dda2c6225a284aa **ADVIL** 1199 9 5d9b4f591dda2c6225a284aa **ADVIL** 1254 5d9b4f591dda2c6225a284aa 10 **ADVIL** 1189 11 5d9b4f591dda2c6225a284aa **ADVIL** 1181 12 ADVIL 1198 13 559c2234e4b06aca36af13c6 AMERICAN BEAUTY 1960 Total rows: 1000 of 2993 Query complete 00:00:00.067

If this information had been given, there would be better way to create fact and dimension tables from the bridge_item_receipts table further down, which would be useful to answer several other business questions.