

For brands, Below showing the percentage of missing data:

ID	0.000000
BARCODE	0.000000
BRAND_CODE	6.157635
CPG_ID	0.000000
CATEGORY	6.650246
CATEGORY_CODE	7.635468
NAME	0.000000
ROMANCE_TEXT	25.369458
RELATED_BRAND_IDS	59.852217

For users, Below showing the percentage of missing data:

CREATED_DATE	0.000000
BIRTH_DATE	0.000000
GENDER	0.000000
LAST_REWARDS_LOGIN	0.000000
STATE	0.000000
SIGN_UP_PLATFORM	27.439024
SIGN_UP_SOURCE	0.000000
ID	0.000000

On the other hand, for receipt\_items, some information is missing. With below showing the percentage of missing data:

REWARDS_RECEIPT_ID	0.000000
ITEM_INDEX	0.000000
REWARDS_RECEIPT_ITEM_ID	0.000000
DESCRIPTION	0.302739
BARCODE	37.563163
BRAND_CODE	57.020842
QUANTITY_PURCHASED	2.152191
TOTAL_FINAL_PRICE	0.192021
POINTS_EARNED	94.741063
REWARDS_GROUP	82.813276
ORIGINAL_RECEIPT_ITEM_TEXT	0.466456
MODIFY_DATE	0.000000

For receipts, some fields are missing. Below shows the percentage of missing data:

ID	0.000000
STORE_NAME	2.600530
PURCHASE_DATE	2.926304
PURCHASE_TIME	7.006983
DATE_SCANNED	0.000000

TOTAL_SPENT	2.113285
REWARDS_RECEIPT_STATUS	0.000000
USER_ID	0.000000
USER_VIEWED	9.157094
PURCHASED_ITEM_COUNT	2.056628
CREATE_DATE	0.000000
PENDING_DATE	2.058045
MODIFY_DATE	0.002833
FLAGGED_DATE	94.298948
PROCESSED_DATE	100.000000
FINISHED_DATE	8.855399
REJECTED_DATE	93.790456
NEEDS_FETCH_REVIEW	99.539667
DIGITAL_RECEIPT	0.000000
DELETED	98.770556
NON_POINT_EARNING_RECEIPT	12.727865

Overall, while some of the fields aren't essential, BRAND\_CODE from brand data set is the only available option to connect between items and brand, there are 25 out of 406 brands have missing BRAND\_CODE; as a result, we won't be able to connect it with items. On the other hand, 57.021% of brand\_code is missing from item side, so we will not be able to know the brand of many items.

For item which I created a table for is only implicitly mentioned in the provided data. The REWARDS\_RECEIPT\_ITEM\_ID seems like to be a possible candidate for primary key of item table; however, all of them turned out to be unique while ORIGINAL\_RECEIPT\_ITEM\_TEXT gives 174085 out of 358696 total values which is might be better candidate for primary key for item table if we believe item with the same text are the same.

For some of the field such as PURCHASED\_ITEM\_COUNT, we have most of the data, so we can use it for potential customer behavior analysis. Such as the chart below shows the count of each item the customer buys. As a result, in most of purchases, customer will only purchase one of each item, so improving variety might be a better strategy and keeping multiple items.



