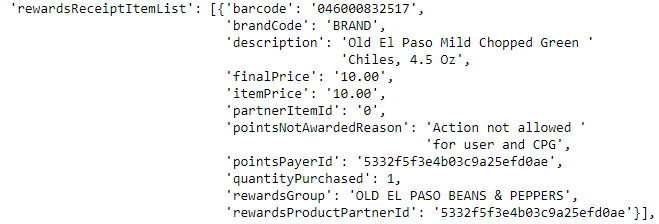
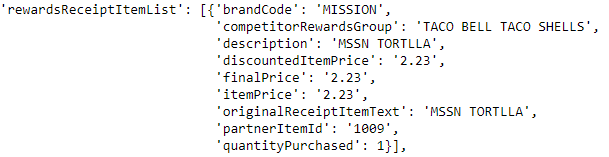
## Basic data quality check

Here are some of the quality issues I found in the limited time I took.

* Null values for a lot of the data points in all the 4 tables.
* Main issue found is there is no consistency in the ReceiptItemList.

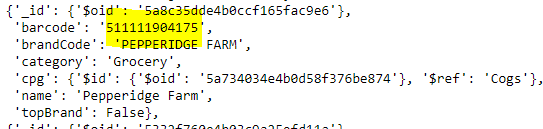


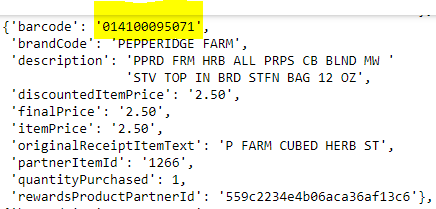


The above two are different data for the ReceiptItemList from two different receipts. Here we can clearly see the inconsistency. The second list is missing crucial value like **barcode** which used to track the brand from the brand table. Not just this some of the barcodes have **4011**, which means the Item value is missing.

Also, the **barcode** for the **brands** are not consistent in the **brands** and **receiptItemList** tables.

Here is an example :





For the brand PEPPERIDGE FARM, we could see that **brands** and **receiptItemList** has different **barcodes.** Which causes a major data quality breach as we are not able to aggregate dat on **barcodes.** Especially when were using the below query to add **brandId** to our **receiptItemList** table –

Select \*

into receiptItemListNew

from (

Select b.[brand\_id] as brandid

,[receiptitem\_id]

,r.[barcode]

,[description]

,[finalPrice]

,[itemPrice]

,[id\_oid] as receipt\_id

from [receiptItemList] r

left join [brands] b

on r.[barcode] = b.[barcode]) t

This is the reason we have **NULL brand IDs** for most rows in the receipyItemList table, which hinders the capability to do joins and derive business insights.

