In [ ]:

```
#This notebook centers on analyzing brand data to detect any possible
# We start by importing the data and standardizing the nested JSON str
```

In [125]:

```
import pandas as pd
import json

# Open the file and read lines
with open('/Users/project/receipts.json', 'r') as file:
    data = file.readlines()
```

In [101]:

```
print(data)
```

```
['{"_id":{"$oid":"5ff1e1eb0a720f0523000575"},"bonusPointsEarned":50
0,"bonusPointsEarnedReason":"Receipt number 2 completed, bonus poin
t schedule DEFAULT (5cefdcacf3693e0b50e83a36)","createDate":{"$dat
e":1609687531000},"dateScanned":{"$date":1609687531000},"finishedDa
te":{"$date":1609687531000},"modifyDate":{"$date":1609687536000},"p
ointsAwardedDate":{"$date":1609687531000},"pointsEarned":"500.0","p
urchaseDate":{"$date":1609632000000},"purchasedItemCount":5,"reward
sReceiptItemList":[{"barcode":"4011","description":"ITEM NOT FOUN
D","finalPrice":"26.00","itemPrice":"26.00","needsFetchReview":fals
e,"partnerItemId":"1","preventTargetGapPoints":true,"quantityPurcha
sed":5,"userFlaggedBarcode":"4011","userFlaggedNewItem":true,"userF
laggedPrice":"26.00","userFlaggedQuantity":5}],"rewardsReceiptStatu
s":"FINISHED","totalSpent":"26.00","userId":"5ff1e1eacfcf6c399c274a
e6"}\n', '{"_id":{"$oid":"5ff1e1bb0a720f052300056b"},"bonusPointsEa
rned":150,"bonusPointsEarnedReason":"Receipt number 5 completed, bo
nus point schedule DEFAULT (5cefdcacf3693e0b50e83a36)","createDat
e":{"$date":1609687483000},"dateScanned":{"$date":1609687483000},"f
inishedDate":{"$date":1609687483000},"modifyDate":{"$date":16096874
88000},"pointsAwardedDate":{"$date":1609687483000},"pointsEarned":"
150.0","purchaseDate":{"$date":1609601083000},"purchasedItemCount":
```

In [ ]:

```
# we an observe that the column 'rewardsReceiptItemList' has a nested
# we split the data, unpack it and merge them together
```

In [126]:

```
import pandas as pd
import json
from pandas import json_normalize

# Initialize empty lists to store the data
main_data = []
rewards_data = []

# Read JSON data from file line by line
with open('/Users/chaitanyavarma/Downloads/receipts.json', 'r') as fil
    for line in file:
        # Load JSON data from each line
        data = json.loads(line)

        # Check if 'rewardsReceiptItemList' key exists
        if 'rewardsReceiptItemList' in data:
            # If key exists, add to rewards_data list
            rewards_data.extend(data['rewardsReceiptItemList'])
            # Remove 'rewardsReceiptItemList' key from data
            del data['rewardsReceiptItemList']

        # Append the remaining data to main_data list
        main_data.append(data)

# Create DataFrame for main data
df_main = pd.json_normalize(main_data)

# Create DataFrame for rewards data
df_rewards = pd.json_normalize(rewards_data)

# Merge DataFrames
receipts_df = pd.merge(df_main, df_rewards, left_index=True, right_inc
```

In [133]:

```
receipts_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1119 entries, 0 to 1118
Data columns (total 48 columns):
 #   Column                                   Non-Null Count  Dtype
---  ------                                   --------------  -----
 0   bonusPointsEarned                        544 non-null    float64
 1   bonusPointsEarnedReason                  544 non-null    object
 2   pointsEarned                             609 non-null    object
 3   purchasedItemCount                       635 non-null    float64
 4   rewardsReceiptStatus                     1119 non-null   object
 5   totalSpent                               684 non-null    object
 6   userId                                   1119 non-null   object
 7   receiptId                                1119 non-null   object
 8   createDate                               1119 non-null   datetime64[n
s]
 9   dateScanned                        1119 non-null   datetime64[n
s]
 10  finishedDate                       568 non-null    datetime64[n
s]
 11  modifyDate                         1119 non-null   datetime64[n
s]
 12  pointsAwardedDate                  537 non-null    datetime64[n
s]
 13  purchaseDate                       671 non-null    datetime64[n
s]
 14  barcode                            775 non-null    object
 15  description                        923 non-null    object
 16  finalPrice                         1014 non-null   object
 17  itemPrice                          1014 non-null   object
 18  needsFetchReview                   208 non-null    object
 19  partnerItemId                      1119 non-null   object
 20  preventTargetGapPoints             200 non-null    object
 21  quantityPurchased                  1014 non-null   float64
 22  userFlaggedBarcode                 200 non-null    object
 23  userFlaggedNewItem                 194 non-null    object
 24  userFlaggedPrice                   178 non-null    object
 25  userFlaggedQuantity                178 non-null    float64
 26  needsFetchReviewReason             133 non-null    object
 27  pointsNotAwardedReason             32 non-null     object
 28  pointsPayerId                      318 non-null    object
 29  rewardsGroup                       441 non-null    object
 30  rewardsProductPartnerId            585 non-null    object
 31  userFlaggedDescription             124 non-null    object
 32  originalMetaBriteBarcode           12 non-null     object
 33  originalMetaBriteDescription       6 non-null      object
 34  brandCode                          331 non-null    object
 35  competitorRewardsGroup             31 non-null     object
 36  discountedItemPrice                496 non-null    object
 37  originalReceiptItemText            491 non-null    object
 38  itemNumber                         5 non-null      object
 39  originalMetaBriteQuantityPurchased  10 non-null     float64
 40  pointsEarned_receiptItem           286 non-null    object
 41  targetPrice                        231 non-null    object
 42  competitiveProduct                 138 non-null    object
 43  originalFinalPrice                 6 non-null      object
 44  originalMetaBriteItemPrice         6 non-null      object
 45  deleted                            3 non-null      object
 46  priceAfterCoupon                   27 non-null     object
 47  metabriteCampaignId                62 non-null     object
dtypes: datetime64[ns](6), float64(5), object(37)
memory usage: 428.4+ KB
```

In [ ]:

```
# we can see that the data has been flatened out.
```

In [128]:

```
# renaming the id field for convinenece
receipts_df.rename(columns={'_id.$oid':'receiptId'}, inplace=True)
```

In [129]:

```
# printing the data all the fullest

pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

In [130]:

```
# converting the dates data into a correct format

receipts_df['dateScanned.$date']   = pd.to_datetime(receipts_df['dateS
receipts_df['createDate.$date']    = pd.to_datetime(receipts_df['creat
receipts_df['finishedDate.$date']   = pd.to_datetime(receipts_df['fini
receipts_df['modifyDate.$date']     = pd.to_datetime(receipts_df['modi
receipts_df['pointsAwardedDate.$date'] = pd.to_datetime(receipts_df['p
receipts_df['purchaseDate.$date']   = pd.to_datetime(receipts_df['purc
```

In [131]:

```
# renaming the columns

receipts_df.rename(columns={'_id.$oid':'receipts_id'}, inplace=True)
receipts_df.rename(columns={'dateScanned.$date':'dateScanned'}, inplac
receipts_df.rename(columns={'createDate.$date':'createDate'}, inplace=
receipts_df.rename(columns={'finishedDate.$date':'finishedDate'}, inpl
receipts_df.rename(columns={'modifyDate.$date':'modifyDate'}, inplace=
receipts_df.rename(columns={'pointsAwardedDate.$date':'pointsAwardedDa
receipts_df.rename(columns={'purchaseDate.$date':'purchaseDate'}, inpl
```

In [ ]:

```
# splitting the data into two dataframes. This will help in creating a

df_rewardreceipts = pd.DataFrame(receipts_df, columns = ['userId','rec
In [114]: df_rewardreceipts.head(3)
```

Out[114]:

	userId	receiptId	barcode	description	finalPrice	i
0	5ff1e1eacfcf6c399c274ae6	5ff1e1eb0a720f0523000575	4011	ITEM NOT FOUND	26.00	
1	5ff1e194b6a9d73a3a9f1052	5ff1e1bb0a720f052300056b	4011	ITEM NOT FOUND	1	
2	5ff1e1f1cfcf6c399c274b0b	5ff1e1f10a720f052300057a	028400642255	DORITOS TORTILLA CHIP SPICY SWEET CHILI REDUCE...	10.00	

In [115]:

```
columns_to_delete = [

    'description',
    'finalPrice',
    'itemPrice',
    'needsFetchReview',
    'partnerItemId',
    'preventTargetGapPoints',
    'quantityPurchased',
    'userFlaggedBarcode',
    'userFlaggedNewItem',
    'userFlaggedPrice',
    'userFlaggedQuantity',
    'needsFetchReviewReason',
    'pointsNotAwardedReason',
    'pointsPayerId',
    'rewardsGroup',
    'rewardsProductPartnerId',
    'userFlaggedDescription',
    'originalMetaBriteBarcode',
    'originalMetaBriteDescription',
    'competitorRewardsGroup',
    'discountedItemPrice',
    'originalReceiptItemText',
    'itemNumber',
    'originalMetaBriteQuantityPurchased',
    'pointsEarned',
    'targetPrice',
    'competitiveProduct',
    'originalFinalPrice',
    'originalMetaBriteItemPrice',
    'deleted',
    'priceAfterCoupon',
    'metabriteCampaignId']


receipts_df = receipts_df.drop(columns=columns_to_delete)
```

In [ ]:

```
# eliminating the records from the original dataframe as we have creat
```

In [134]:

```
receipts_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1119 entries, 0 to 1118
Data columns (total 48 columns):
 #   Column                                    Non-Null Count  Dtype
---  ------                                    --------------  -----
```

```
 0   bonusPointsEarned                   544 non-null    float64
 1   bonusPointsEarnedReason             544 non-null    object
 2   pointsEarned                        609 non-null    object
 3   purchasedItemCount                  635 non-null    float64
 4   rewardsReceiptStatus                1119 non-null   object
 5   totalSpent                          684 non-null    object
 6   userId                              1119 non-null   object
 7   receiptId                           1119 non-null   object
 8   createDate                          1119 non-null   datetime64[n
s]
 9   dateScanned                         1119 non-null   datetime64[n
s]
 10  finishedDate                        568 non-null    datetime64[n
s]
 11  modifyDate                          1119 non-null   datetime64[n
s]
 12  pointsAwardedDate                   537 non-null    datetime64[n
s]
 13  purchaseDate                        671 non-null    datetime64[n
s]
 14  barcode                             775 non-null    object
 15  description                         923 non-null    object
 16  finalPrice                          1014 non-null   object
 17  itemPrice                           1014 non-null   object
 18  needsFetchReview                    208 non-null    object
 19  partnerItemId                       1119 non-null   object
 20  preventTargetGapPoints              200 non-null    object
 21  quantityPurchased                   1014 non-null   float64
 22  userFlaggedBarcode                  200 non-null    object
 23  userFlaggedNewItem                  194 non-null    object
 24  userFlaggedPrice                    178 non-null    object
 25  userFlaggedQuantity                 178 non-null    float64
 26  needsFetchReviewReason              133 non-null    object
 27  pointsNotAwardedReason              32 non-null     object
 28  pointsPayerId                       318 non-null    object
 29  rewardsGroup                        441 non-null    object
 30  rewardsProductPartnerId             585 non-null    object
 31  userFlaggedDescription              124 non-null    object
 32  originalMetaBriteBarcode            12 non-null     object
 33  originalMetaBriteDescription        6 non-null      object
 34  brandCode                           331 non-null    object
 35  competitorRewardsGroup              31 non-null     object
 36  discountedItemPrice                 496 non-null    object
 37  originalReceiptItemText             491 non-null    object
 38  itemNumber                          5 non-null      object
 39  originalMetaBriteQuantityPurchased  10 non-null     float64
 40  pointsEarned_receiptItem            286 non-null    object
 41  targetPrice                         231 non-null    object
 42  competitiveProduct                  138 non-null    object
 43  originalFinalPrice                  6 non-null      object
 44  originalMetaBriteItemPrice         6 non-null      object
 45  deleted                            3 non-null      object
 46  priceAfterCoupon                   27 non-null     object
 47  metabriteCampaignId                62 non-null     object
dtypes: datetime64[ns](6), float64(5), object(37)
memory usage: 428.4+ KB
```

In [117]:

```
# checking for duplicate values in 'receipts' dataset

duplicate_rows = receipts_df[receipts_df.duplicated()]
duplicate_rows.count()
```

Out[117]:

```
bonusPointsEarned          0
bonusPointsEarnedReason    0
purchasedItemCount         0
rewardsReceiptStatus       0
totalSpent                 0
userId                     0
receiptId                  0
createDate                 0
dateScanned                0
finishedDate               0
modifyDate                 0
pointsAwardedDate          0
purchaseDate               0
barcode                    0
brandCode                  0
pointsEarned_receiptItem   0
dtype: int64
```

In [119]:

```
# checking for duplicate values in 'rewardreceipts' dataset

duplicate_rows = df_rewardreceipts[df_rewardreceipts.duplicated()]
duplicate_rows.count()
```

Out[119]:

```
userId                                0
receiptId                             0
barcode                               0
description                           0
finalPrice                            0
itemPrice                             0
needsFetchReview                      0
partnerItemId                         0
preventTargetGapPoints                0
quantityPurchased                     0
userFlaggedBarcode                    0
userFlaggedNewItem                    0
userFlaggedPrice                      0
userFlaggedQuantity                   0
needsFetchReviewReason                0
pointsNotAwardedReason                0
pointsPayerId                         0
rewardsGroup                          0
rewardsProductPartnerId               0
userFlaggedDescription                0
originalMetaBriteBarcode              0
originalMetaBriteDescription          0
brandCode                             0
competitorRewardsGroup                0
discountedItemPrice                   0
originalReceiptItemText               0
itemNumber                            0
originalMetaBriteQuantityPurchased    0
pointsEarned                          0
targetPrice                           0
competitiveProduct                    0
originalFinalPrice                    0
originalMetaBriteItemPrice            0
deleted                               0
priceAfterCoupon                      0
metabriteCampaignId                   0
dtype: int64
```

In [ ]:

```
# There are absolutely no duplicated records in both the dataset
```

In [120]:

```
# checking for null values in 'rewardreceipts' dataset

df_rewardreceipts.isnull().sum()
```

Out[120]:

```
userId                                       0
receiptId                                    0
barcode                                    344
description                                196
finalPrice                                 105
itemPrice                                  105
needsFetchReview                           911
partnerItemId                                0
preventTargetGapPoints                     919
quantityPurchased                          105
userFlaggedBarcode                         919
userFlaggedNewItem                         925
userFlaggedPrice                           941
userFlaggedQuantity                        941
needsFetchReviewReason                     986
pointsNotAwardedReason                    1087
pointsPayerId                              801
rewardsGroup                               678
rewardsProductPartnerId                    534
userFlaggedDescription                     995
originalMetaBriteBarcode                  1107
originalMetaBriteDescription              1113
brandCode                                  788
competitorRewardsGroup                    1088
discountedItemPrice                        623
originalReceiptItemText                    628
itemNumber                                1114
originalMetaBriteQuantityPurchased        1109
pointsEarned                               510
targetPrice                                888
competitiveProduct                         981
originalFinalPrice                        1113
originalMetaBriteItemPrice                1113
deleted                                   1116
priceAfterCoupon                          1092
metabriteCampaignId                       1057
dtype: int64
```

In [121]:

```
# checking for null values in 'Receipts' dataset

receipts_df.isnull().sum()
```

Out[121]:

```
bonusPointsEarned           575
bonusPointsEarnedReason     575
purchasedItemCount          484
rewardsReceiptStatus          0
totalSpent                  435
userId                        0
receiptId                     0
createDate                    0
dateScanned                   0
finishedDate                551
modifyDate                    0
pointsAwardedDate           582
purchaseDate                448
barcode                     344
brandCode                   788
pointsEarned_receiptItem    833
dtype: int64
```

In [124]:

```
# cheking in terms of percentage

percentage_null_values = (df_rewardreceipts.isnull().sum() / len(df_re
print(percentage_null_values)
```

```
userId                                 0.000000
receiptId                              0.000000
barcode                               30.741734
description                           17.515639
finalPrice                             9.383378
itemPrice                              9.383378
needsFetchReview                      81.411975
partnerItemId                          0.000000
preventTargetGapPoints                82.126899
quantityPurchased                      9.383378
userFlaggedBarcode                    82.126899
userFlaggedNewItem                    82.663092
userFlaggedPrice                      84.092940
userFlaggedQuantity                   84.092940
needsFetchReviewReason                88.114388
pointsNotAwardedReason                97.140304
pointsPayerId                         71.581769
rewardsGroup                          60.589812
rewardsProductPartnerId               47.721180
userFlaggedDescription                88.918677
originalMetaBriteBarcode              98.927614
originalMetaBriteDescription          99.463807
brandCode                             70.420018
competitorRewardsGroup                97.229669
discountedItemPrice                   55.674710
originalReceiptItemText               56.121537
itemNumber                            99.553172
originalMetaBriteQuantityPurchased    99.106345
pointsEarned                          45.576408
targetPrice                           79.356568
competitiveProduct                    87.667560
originalFinalPrice                    99.463807
originalMetaBriteItemPrice            99.463807
deleted                               99.731903
priceAfterCoupon                      97.587131
metabriteCampaignId                   94.459339
dtype: float64
```

In [123]:

```
percentage_null_values = (receipts_df.isnull().sum() / len(receipts_df
print(percentage_null_values)
```

```
bonusPointsEarned             51.385165
bonusPointsEarnedReason       51.385165
purchasedItemCount            43.252904
rewardsReceiptStatus           0.000000
totalSpent                    38.873995
userId                         0.000000
receiptId                      0.000000
createDate                     0.000000
dateScanned                    0.000000
finishedDate                  49.240393
modifyDate                     0.000000
pointsAwardedDate             52.010724
purchaseDate                  40.035746
barcode                       30.741734
brandCode                     70.420018
pointsEarned_receiptItem      74.441466
dtype: float64
```

In [ ]:

```
# There are one to many null values in both the datasets. There are ev
```

In [ ]:

```
# Issues or Anamolies present in the receipts and rewardreceipts datas

# The initial receipts data contains deeply nested json values.
# There are no duplicated values present in both the datasets.
# There are a lot of null values that are present in both the datasets
# There are columns with more than 60-90 % of them that are empty
```