1 Importing library and data

In [9]:

```
import pandas as pd
it = pd.read_csv('item.csv')
```

2 Check the data type

In [10]:

```
it.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6941 entries, 0 to 6940
Data columns (total 35 columns):
 #   Column                              Non-Null Count  Dtype  
---  ------                              --------------  -----  
 0   OID                                 6941 non-null   object 
 1   barcode                             3090 non-null   object 
 2   description                         6560 non-null   object 
 3   finalPrice                          6767 non-null   float64
 4   itemPrice                           6767 non-null   float64
 5   needsFetchReview                    813 non-null    object 
 6   partnerItemId                       6941 non-null   int64  
 7   preventTargetGapPoints              358 non-null    object 
 8   quantityPurchased                   6767 non-null   float64
 9   userFlaggedBarcode                  337 non-null    float64
 10  userFlaggedNewItem                  323 non-null    object 
 11  userFlaggedPrice                    299 non-null    float64
 12  userFlaggedQuantity                 299 non-null    float64
 13  needsFetchReviewReason              219 non-null    object 
 14  pointsNotAwardedReason              340 non-null    object 
 15  pointsPayerId                       1267 non-null   object 
 16  rewardsGroup                        1731 non-null   object 
 17  rewardsProductPartnerId             2269 non-null   object 
 18  userFlaggedDescription              154 non-null    object 
 19  originalMetaBriteBarcode            24 non-null     float64
 20  originalMetaBriteDescription        10 non-null     object 
 21  brandCode                           2600 non-null   object 
 22  competitorRewardsGroup              275 non-null    object 
 23  discountedItemPrice                 5769 non-null   float64
 24  originalReceiptItemText             5760 non-null   object 
 25  itemNumber                          153 non-null    float64
 26  originalMetaBriteQuantityPurchased  15 non-null     float64
 27  pointsEarned                        927 non-null    float64
 28  targetPrice                         378 non-null    float64
 29  competitiveProduct                  645 non-null    object 
 30  originalFinalPrice                  9 non-null      float64
 31  originalMetaBriteItemPrice          9 non-null      float64
 32  deleted                             9 non-null      object 
 33  priceAfterCoupon                    956 non-null    float64
 34  metabriteCampaignId                 863 non-null    object 
dtypes: float64(15), int64(1), object(19)
memory usage: 1.9+ MB
```

In [11]:

```
it.head()
```

Out[11]:

	OID	barcode	description	finalPrice	itemPrice	needsFetchRevi
0	5ff1e1eb0a720f0523000575	4011	ITEM NOT FOUND	26.0	26.0	Fa
1	5ff1e1bb0a720f052300056b	4011	ITEM NOT FOUND	1.0	1.0	N
2	5ff1e1bb0a720f052300056b	028400642255	DORITOS TORTILLA CHIP SPICY SWEET CHILI REDUCE...	10.0	10.0	Tr
3	5ff1e1f10a720f052300057a	NaN	NaN	NaN	NaN	Fa
4	5ff1e1ee0a7214ada100056f	4011	ITEM NOT FOUND	28.0	28.0	Fa

5 rows × 35 columns

In [12]:

```
# Check the type of each feature in the OID table
print(it.dtypes)
```

```
OID                                  object
barcode                              object
description                          object
finalPrice                          float64
itemPrice                           float64
needsFetchReview                     object
partnerItemId                         int64
preventTargetGapPoints               object
quantityPurchased                   float64
userFlaggedBarcode                  float64
userFlaggedNewItem                   object
userFlaggedPrice                    float64
userFlaggedQuantity                 float64
needsFetchReviewReason               object
pointsNotAwardedReason               object
pointsPayerId                        object
rewardsGroup                         object
rewardsProductPartnerId              object
userFlaggedDescription               object
originalMetaBriteBarcode            float64
originalMetaBriteDescription         object
brandCode                            object
competitorRewardsGroup               object
discountedItemPrice                 float64
originalReceiptItemText              object
itemNumber                          float64
originalMetaBriteQuantityPurchased  float64
pointsEarned                        float64
targetPrice                         float64
competitiveProduct                   object
originalFinalPrice                  float64
originalMetaBriteItemPrice          float64
deleted                              object
priceAfterCoupon                    float64
metabriteCampaignId                  object
dtype: object
```

3 The number of null value of each column

In [13]:

```
# Count the number of null value of each column
num_missing = it.isnull().sum()
print(num_missing)
```

```
OID                                    0
barcode                             3851
description                          381
finalPrice                           174
itemPrice                            174
needsFetchReview                    6128
partnerItemId                          0
preventTargetGapPoints              6583
quantityPurchased                    174
userFlaggedBarcode                  6604
userFlaggedNewItem                  6618
userFlaggedPrice                    6642
userFlaggedQuantity                 6642
needsFetchReviewReason              6722
pointsNotAwardedReason              6601
pointsPayerId                       5674
rewardsGroup                        5210
rewardsProductPartnerId             4672
userFlaggedDescription              6787
originalMetaBriteBarcode            6917
originalMetaBriteDescription        6931
brandCode                           4341
competitorRewardsGroup              6666
discountedItemPrice                 1172
originalReceiptItemText             1181
itemNumber                          6788
originalMetaBriteQuantityPurchased  6926
pointsEarned                        6014
targetPrice                         6563
competitiveProduct                  6296
originalFinalPrice                  6932
originalMetaBriteItemPrice          6932
deleted                             6932
priceAfterCoupon                    5985
metabriteCampaignId                 6078
dtype: int64
```

In [14]:

```
# Calculate the percentage of null values in each column
null_percentages = (it.isnull().sum() / len(it)) * 100

print(null_percentages)
```

```
OID                                   0.000000
barcode                              55.481919
description                           5.489123
finalPrice                            2.506843
itemPrice                             2.506843
needsFetchReview                     88.286990
partnerItemId                         0.000000
preventTargetGapPoints               94.842242
quantityPurchased                     2.506843
userFlaggedBarcode                   95.144792
userFlaggedNewItem                   95.346492
userFlaggedPrice                     95.692263
userFlaggedQuantity                  95.692263
needsFetchReviewReason               96.844835
pointsNotAwardedReason               95.101570
pointsPayerId                        81.746146
rewardsGroup                         75.061230
rewardsProductPartnerId              67.310186
userFlaggedDescription               97.781300
originalMetaBriteBarcode             99.654228
originalMetaBriteDescription         99.855929
brandCode                            62.541421
competitorRewardsGroup               96.038035
discountedItemPrice                  16.885175
originalReceiptItemText              17.014839
itemNumber                           97.795707
originalMetaBriteQuantityPurchased   99.783893
pointsEarned                         86.644576
targetPrice                          94.554099
competitiveProduct                   90.707391
originalFinalPrice                   99.870336
originalMetaBriteItemPrice           99.870336
deleted                              99.870336
priceAfterCoupon                     86.226768
metabriteCampaignId                  87.566633
dtype: float64
```

4 Percentage of duplicate value in 'OID' column

In [18]:

```
# Check the number of duplicate value in 'OID' column
duplicate_var = it['OID'].duplicated().sum()
print(duplicate_var)
```

```
6262
```

In [19]:

```
# Calculate the percentage of duplicate value in 'userId' column
duplicates_num = (it['OID'].count() - it['OID'].nunique())
percentage_dup = (duplicates_num / it['OID'].count()) * 100
print("The percentage of dupplicate value in 'OID' column is: ", percentage_dup)
```

```
The percentage of dupplicate value in 'OID' column is:  90.21754790376026
```

In [ ]:

```

```