1 Importing library and data

In [15]:

```
import pandas as pd
br = pd.read_csv('brands.csv')
```

2 Check the data type

In [16]:

```
br.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1167 entries, 0 to 1166
Data columns (total 9 columns):
 #   Column        Non-Null Count  Dtype  
---  ------        --------------  -----  
 0   MID           1167 non-null   object 
 1   barcode       1167 non-null   int64  
 2   CATEGORY      1012 non-null   object 
 3   CATEGORYCODE  517 non-null    object 
 4   REF           1167 non-null   object 
 5   PARTNERID     1167 non-null   object 
 6   NAME          1167 non-null   object 
 7   TOPBRAND      555 non-null    float64
 8   BRANDCODE     898 non-null    object 
dtypes: float64(1), int64(1), object(7)
memory usage: 82.2+ KB
```

In [17]:

```
br.head()
```

Out[17]:

	MID	barcode	CATEGORY	CATEGORYCODE	REF	
0	601ac115be37ce2ead437551	511111019862	Baking	BAKING	Cogs	601ac
1	601c5460be37ce2ead43755f	511111519928	Beverages	BEVERAGES	Cogs	5332
2	601ac142be37ce2ead43755d	511111819905	Baking	BAKING	Cogs	601ac
3	601ac142be37ce2ead43755a	511111519874	Baking	BAKING	Cogs	601ac
4	601ac142be37ce2ead43755e	511111319917	Candy & Sweets	CANDY_AND_SWEETS	Cogs	5332I

In [18]:

```
# Check the type of each feature in the brands table
print(br.dtypes)
```

```
MID              object
barcode           int64
CATEGORY         object
CATEGORYCODE     object
REF              object
PARTNERID        object
NAME             object
TOPBRAND        float64
BRANDCODE        object
dtype: object
```

3 The number of null value of each column

In [19]:

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

```
MID                0
barcode            0
CATEGORY         155
CATEGORYCODE     650
REF                0
PARTNERID          0
NAME               0
TOPBRAND         612
BRANDCODE        269
dtype: int64
```

In [20]:

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

print(null_percentages)
```

```
MID               0.000000
barcode           0.000000
CATEGORY         13.281919
CATEGORYCODE     55.698372
REF               0.000000
PARTNERID         0.000000
NAME              0.000000
TOPBRAND         52.442159
BRANDCODE        23.050557
dtype: float64
```

4 Percentage of duplicate value in 'BRANDCODE' column

In [21]:

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

```
270
```

In [22]:

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

```
The percentage of dupplicate value in 'BRANDCODE' column is:  0.2227171492
2048996
```

In [ ]:

```

```