Import Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Load the input files

```
product = pd.read_csv('PRODUCTS_TAKEHOME.csv')
product.head()
```

	CATEGORY_1	CATEGORY_2	CATEGORY_3	CATEGORY_4	MANUFACTURER	BRAND	BARCODE
0	Health & Wellness	Sexual Health	Conductivity Gels & Lotions	NaN	NaN	NaN	7.964944e+11
1	Snacks	Puffed Snacks	Cheese Curls & Puffs	NaN	NaN	NaN	2.327801e+10
2	Health & Wellness	Hair Care	Hair Care Accessories	NaN	PLACEHOLDER MANUFACTURER	ELECSOP	4.618178e+11
	Health &						

```
product.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116005 entries, 0 to 116004
Data columns (total 7 columns):
 #   Column        Non-Null Count   Dtype
---  ------        --------------   -----
 0   CATEGORY_1    115989 non-null  object
 1   CATEGORY_2    115813 non-null  object
 2   CATEGORY_3    107822 non-null  object
 3   CATEGORY_4    9268 non-null    object
 4   MANUFACTURER  84772 non-null   object
 5   BRAND         84772 non-null   object
 6   BARCODE       115448 non-null  float64
dtypes: float64(1), object(6)
memory usage: 6.2+ MB
```

Converting the data types

```
product['BARCODE'] = product['BARCODE'].fillna(0).astype(int)
product.head()
```

	CATEGORY_1	CATEGORY_2	CATEGORY_3	CATEGORY_4	MANUFACTURER	BRAND	BARCODE
0	Health & Wellness	Sexual Health	Conductivity Gels & Lotions	NaN	NaN	NaN	796494407820
1	Snacks	Puffed Snacks	Cheese Curls & Puffs	NaN	NaN	NaN	23278011028
2	Health & Wellness	Hair Care	Hair Care Accessories	NaN	PLACEHOLDER MANUFACTURER	ELECSOP	461817824225
	Health &						

Converting the data type for BARCODE into integer value.

Missing data count

```
product.isnull().sum()
```

	0
CATEGORY_1	16
CATEGORY_2	192
CATEGORY_3	8183
CATEGORY_4	106737
MANUFACTURER	31233
BRAND	31233
BARCODE	0

dtype: int64

Percentage of missing data in each column

```
product.isnull().sum() / len(product) * 100
```

	0
CATEGORY_1	0.013793
CATEGORY_2	0.165510
CATEGORY_3	7.054006
CATEGORY_4	92.010689
MANUFACTURER	26.923839
BRAND	26.923839
BARCODE	0.000000

dtype: float64

We have significant amount of data missing in CATEGORY_4 column around 92 percent and 26 percent of data missing in MANUFACTURER and BRAND columns.

Checking for duplicates

```
product.duplicated().sum()
```

```
5
```

Around 215 duplicates rows were detected in the product dataset.

Unique values in columns

```
product['BARCODE'].is_unique
```

```
False
```

All the null values which are represented as 0's in the column are to be removed to make barcode as primary key for this dataset.

```
product['CATEGORY_1'].unique()
```

```
array(['Health & Wellness', 'Snacks', 'Beverages', 'Pantry', 'Alcohol',
       'Apparel & Accessories', 'Restaurant', 'Needs Review', 'Dairy',
       'Home & Garden', nan, 'Household Supplies', 'Meat & Seafood',
       'Deli & Bakery', 'Sporting Goods', 'Produce', 'Office & School',
       'Frozen', 'Arts & Entertainment', 'Animals & Pet Supplies',
       'Electronics', 'Beauty', 'Toys & Games', 'Mature',
       'Vehicles & Parts', 'Baby & Toddler', 'Luggage & Bags', 'Media'],
      dtype=object)
```

```
product['MANUFACTURER'].unique()
```

```
array([nan, 'PLACEHOLDER MANUFACTURER', 'COLGATE-PALMOLIVE', ...,
       'VIDETTE INC', 'SCRUB-IT', 'OUTDOOR PRODUCT INNOVATIONS, INC.'],
      dtype=object)
```

```
product['BRAND'].unique()
```

```
array([nan, 'ELECSOP', 'COLGATE', ..., 'SHULEMIN', 'RHINO BLINDS',
       'GATEWAY'], dtype=object)
```

Several issues were identified in the PRODUCT dataset:

1) A significant amount of data is missing in the CATEGORY_4 (92%), MANUFACTURER (26%), and BRAND (26%) columns.

2) There are 215 duplicate rows in the dataset.

3) The BARCODE column is in float type, making it difficult to use as a categorical variable. Additionally, it contains null values, which should either be removed or updated with the correct information to enable the column to serve as the primary key for the dataset.

Challenges faced:

The presence of over 26% missing data in the BRAND and MANUFACTURER columns significantly reduces their effectiveness in providing valuable insights.

All the above data issues are to be resolved before the dataset is used for further steps of the process.