Section 3: Identifying Famous Brands and Categories

In this lesson, the most famous brands and categories are fetched from our dataset.

We'll cover the following

The most common problem faced by any business is inventory management. Sometimes business owners either have too much of a product that is not being sold or too little of a product whose demand is very high. This can cause a substantial loss to a company's profits and reputation. For more information on this problem, refer here.

If we somehow know what products from which brands and categories are selling the most in the market, then inventory management can be optimized to some level. Here, products from which brand and category were bought the most will be determined.

Top brands

First, the data will be read and the `event_time` column will be converted to DateTime format. Then, the following steps will be performed to obtain the top brands.

```
import pandas as pd

df = pd.read_csv("2019-Oct.csv")
df['event_time'] = pd.to_datetime(df.event_time)

# Get rows where products are purchased
purchase = df[df['event_type'] == 'purchase']

# Group the DataFrame on brands
top_brands = purchase.groupby('brand')

# Get number of products bought by computing length of each grouped brand
top_brands = top_brands['brand'].agg([len])
```

```
# Sort the result on obtained length in descending order
top_brands.sort_values('len', ascending = False, inplace = True)


print(top_brands)
```

brand	len
samsung	172896
apple	142873
xiaomi	56616
huawei	23501
lucente	11578
...	...
plantex	1
playmates	1
libbey	1
alser	1
borasco	1

1983 rows × 1 columns

On **line** 7, the rows whose `event_type` is equal to `purchase` are selected as we are trying to access results for purchased products.

On **line 10**, the `purchase` DataFrame is grouped on the `brand` column. We need the information on how many products are bought for each brand against it. Grouping on the `brand` column separates all the brands and their relevant information in their own groups.

On **line 13**, the length for each group is calculated using the `agg()` function. As in each group, the brand names are listed with their products. Finding the length gives us how many products are bought against each brand.

On **line 16**, the result is sorted in descending order of length so we can get which brand has most sold products on the top and the brand with least sold products in

the end.

According to this, `Samsung` is the most famous brand whose products are being bought in excessive quantities.

Top categories

The same steps as above will be performed here, but instead of the `brand` column, the `category_code` column will be used.

```
import pandas as pd

df = pd.read_csv("2019-Oct.csv")
df['event_time'] = pd.to_datetime(df.event_time)

# Get rows where products are purchased
purchase = df[df['event_type'] == 'purchase']

# Group the DataFrame on category_code
top_catg = purchase.groupby('category_code')

# Get number of products bought by computing length of each grouped category_code
top_catg = top_catg['category_code'].agg([len])

# Sort the result on obtained length in descending order
top_catg.sort_values('len', ascending = False, inplace = True)

print(top_catg)
```

category_code	len
electronics.smartphone	338018
electronics.audio.headphone	30503
electronics.video.tv	21565
electronics.clocks	17906
appliances.kitchen.washer	16148
...	...
apparel.skirt	4
apparel.shorts	2
apparel.jumper	2
construction.tools.soldering	1
apparel.jacket	1

121 rows × 1 columns

The same technique and codes to find the top brands are used to get the top categories. Only the `brand` column is replaced with the `category_code` column.

According to the above output, the smartphone category is the most famous among others. The difference in the number of products bought for other categories is clearly visible.

The hypothesis

According to the above analysis, the top brands all include mobile and mobile accessory companies. The top category is the smartphone category, which has over **300,000**+ sales, and the other categories don't even come close to this number. It can be concluded that all products that come under the smartphone category should be in abundance in the inventory with only the top five or six top brands.

In the next lesson, the recency-frequency-monetary (RFM) analysis is performed

on the dataset.