Data Science Assignment-2

Problem statement:
Using ML/DL techniques, match similar products from the Flipkart dataset with the Amazon dataset. Once similar products are matched, display the retail price from FK and AMZ side by side. Please explore as many techniques as possible before choosing the final technique. You may either display the final result in single table format OR You may create a simple form where we input the product name and the output of prices of the product from both websites are displayed.

In [1]:
```
import pandas as pd
import numpy as np
from pandasql import sqldf
```

In [2]:
```
az_data=pd.read_csv('amz_com-ecommerce_sample.csv',encoding='unicode_escape')
fp_data=pd.read_csv('flipkart_com-ecommerce_sample.csv',encoding='unicode_escape')
```

Amazon Data Preprocessing

In [3]:
```
az_data.head()
```

Out[3]:

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_
0	c2d766ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2FF9KEDEFGF	
1	7f7036a6d550aaa89d34c77bd39a5e48	2016-03-25 22:59:23 +0000	http://www.flipkart.com/fabhomedecor-fabric-do...	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	SBEEH3QGU7MFYJFY	
2	f449ec65dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	SHOEH4GRSUBJGZXE	
3	0973b37acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2F6HUZMQ6SJ	
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	2016-03-25 22:59:23 +0000	http://www.flipkart.com/sicons-all-purpose-arn...	Sicons All Purpose Arnica Dog Shampoo	["Pet Supplies >> Grooming >> Skin & Coat Care...	PSOEH3ZYDMSYARJ5	

In [4]:
```
az_data.shape
```

Out[4]:
```
(20000, 15)
```

In [5]:
```
az_data.isnull().sum()
```

Out[5]:
```
uniq_id                       0
crawl_timestamp               0
product_url                   0
product_name                  0
product_category_tree         0
pid                           0
retail_price                  0
discounted_price              0
image                         3
is_FK_Advantage_product       0
description                   2
product_rating                0
overall_rating                0
brand                      5864
product_specifications       14
dtype: int64
```

In [6]:

```
az_data = az_data[["uniq_id","product_name","retail_price", "discounted_price"]]
az_data.rename(columns = {'product_name':'amazon_product_name', 'retail_price':'amazon_retail_price',
                          'discounted_price':'amazon_discounted_price'}, inplace = True)
az_data.head()
```

Out[6]:

	uniq_id	amazon_product_name	amazon_retail_price	amazon_discounted_price
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	982	438
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32143	29121
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	991	551
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	694	325
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	208	258

In [7]:

```
az_data.shape
```

Out[7]:

```
(20000, 4)
```

Flipkart Data Preprocessing

In [8]:

```
fp_data=pd.read_csv('flipkart_com-ecommerce_sample.csv',encoding='unicode_escape')
fp_data.head()
```

Out[8]:

	uniq_id	crawl_timestamp	product_url	product_name	product_category_tree	pid	retail_
0	c2d766ca982eca8304150849735ffef9	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2FF9KEDEFGF	
1	7f7036a6d550aaa89d34c77bd39a5e48	2016-03-25 22:59:23 +0000	http://www.flipkart.com/fabhomedecor-fabric-do...	FabHomeDecor Fabric Double Sofa Bed	["Furniture >> Living Room Furniture >> Sofa B...	SBEEH3QGU7MFYJFY	32
2	f449ec65dcbc041b6ae5e6a32717d01b	2016-03-25 22:59:23 +0000	http://www.flipkart.com/aw-bellies/p/itmeh4grg...	AW Bellies	["Footwear >> Women's Footwear >> Ballerinas >...	SHOEH4GRSUBJGZXE	
3	0973b37acd0c664e3de26e97e5571454	2016-03-25 22:59:23 +0000	http://www.flipkart.com/alisha-solid-women-s-c...	Alisha Solid Women's Cycling Shorts	["Clothing >> Women's Clothing >> Lingerie, Sl...	SRTEH2F6HUZMQ6SJ	
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	2016-03-25 22:59:23 +0000	http://www.flipkart.com/sicons-all-purpose-arn...	Sicons All Purpose Arnica Dog Shampoo	["Pet Supplies >> Grooming >> Skin & Coat Care...	PSOEH3ZYDMSYARJ5	

In [9]:

```
fp_data = fp_data[["uniq_id","product_name","retail_price", "discounted_price"]]
fp_data.rename(columns = {'product_name':'flipkart_product_name', 'retail_price':'flipkart_retail_price',
                          'discounted_price':'flipkart_discounted_price'}, inplace = True)
fp_data.head()
```

Out[9]:

	uniq_id	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	999.0	379.0
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32157.0	22646.0
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	999.0	499.0
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	699.0	267.0
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	220.0	210.0

In [10]:

```
fp_data.shape
```

Out[10]:

```
(20000, 4)
```

In [11]:

```
fp_data.isnull().sum()
```

Out[11]:

```
uniq_id                       0
flipkart_product_name         0
flipkart_retail_price        78
flipkart_discounted_price    78
dtype: int64
```

In [12]:

```
null_Values = fp_data[fp_data['flipkart_retail_price'].isna()]
null_Values.head()
```

Out[12]:

	uniq_id	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
12	c29af37837afcaf44b779eca7c19295f	Sicons All Purpose Tea Tree Dog Shampoo	NaN	NaN
21	ea98a65ad1e1b8688eddf89fbc7b3e27	Alisha Solid Women's Cycling Shorts	NaN	NaN
76	64d5d4a258243731dc7bbb1eef49ad74	Eurospa Cotton Terry Face Towel Set	NaN	NaN
812	39ed975091fd6d1d3ef3a7d33b8c4360	Fundoo T Printed Men's Track Suit	NaN	NaN
1318	115ecd52f86ebf7a5509d0959a2caaa0	Techware Microwavable Tea Cups WF13115 - Purpl...	NaN	NaN

In [13]:

```
na=null_Values.index
na
```

Out[13]:

```
Int64Index([   12,    21,    76,   812,  1318,  1734,  2977,  3233,  3404,
             3889,  4590,  5644,  5696,  5967,  6248,  6271,  6601,  6874,
             7396,  7646,  7725,  8027,  8040,  8117,  8176,  8187,  8190,
             8267,  8383,  8448,  8500,  8918,  9069, 10117, 10135, 10272,
            10669, 10745, 11072, 11118, 11319, 11337, 11379, 11383, 11393,
            11420, 11432, 11499, 11567, 11682, 11961, 11962, 11981, 12052,
            12120, 12226, 12274, 12336, 12395, 12396, 12401, 12433, 12541,
            12572, 12690, 12734, 12933, 13123, 13354, 16143, 16296, 16312,
            16487, 16762, 17634, 19543, 19599, 19622],
           dtype='int64')
```

Drop null values from flipkart data as well as amazon data to maintain the accuracy after combining the data.

In [14]:

```
az_data.drop([12,21,76,812,1318,1734,2977,3233,3404,
             3889,  4590,  5644,  5696,  5967,  6248,  6271,  6601,  6874,
             7396,  7646,  7725,  8027,  8040,  8117,  8176,  8187,  8190,
             8267,  8383,  8448,  8500,  8918,  9069, 10117, 10135, 10272,
            10669, 10745, 11072, 11118, 11319, 11337, 11379, 11383, 11393,
            11420, 11432, 11499, 11567, 11682, 11961, 11962, 11981, 12052,
            12120, 12226, 12274, 12336, 12395, 12396, 12401, 12433, 12541,
            12572, 12690, 12734, 12933, 13123, 13354, 16143, 16296, 16312,
            16487, 16762, 17634, 19543, 19599, 19622],axis=0,inplace=True)
```

In [15]:

```
fp_data.drop([12,21,76,812,1318,1734,2977,3233,3404,
             3889,  4590,  5644,  5696,  5967,  6248,  6271,  6601,  6874,
             7396,  7646,  7725,  8027,  8040,  8117,  8176,  8187,  8190,
             8267,  8383,  8448,  8500,  8918,  9069, 10117, 10135, 10272,
            10669, 10745, 11072, 11118, 11319, 11337, 11379, 11383, 11393,
            11420, 11432, 11499, 11567, 11682, 11961, 11962, 11981, 12052,
            12120, 12226, 12274, 12336, 12395, 12396, 12401, 12433, 12541,
            12572, 12690, 12734, 12933, 13123, 13354, 16143, 16296, 16312,
            16487, 16762, 17634, 19543, 19599, 19622],axis=0,inplace=True)
```

In [16]:

```
az_data.shape
```

Out[16]:

```
(19922, 4)
```

In [17]:

```
fp_data.shape
```

Out[17]:

```
(19922, 4)
```

Method-1

Concatenate one DataFrame to those of another DataFrame

In [18]:

```
data = pd.concat([az_data,fp_data],axis=1)
```

In [19]:
```
data.head()
```

Out[19]:

	uniq_id	amazon_product_name	amazon_retail_price	amazon_discounted_price	uniq_id	flipkart_produ
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	982	438	c2d766ca982eca8304150849735ffef9	Alisha Solid Cyclir
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32143	29121	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDec Double
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	991	551	f449ec65dcbc041b6ae5e6a32717d01b	A
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	694	325	0973b37acd0c664e3de26e97e5571454	Alisha Solid Cyclir
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	208	258	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Arnica Dog

In [20]:
```
data = data.drop(['uniq_id'],axis = 1)
```

In [21]:
```
data.shape
```

Out[21]:
```
(19922, 6)
```

In [22]:
```
data.head()
```

Out[22]:

	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
0	Alisha Solid Women's Cycling Shorts	982	438	Alisha Solid Women's Cycling Shorts	999.0	379.0
1	FabHomeDecor Fabric Double Sofa Bed	32143	29121	FabHomeDecor Fabric Double Sofa Bed	32157.0	22646.0
2	AW Bellies	991	551	AW Bellies	999.0	499.0
3	Alisha Solid Women's Cycling Shorts	694	325	Alisha Solid Women's Cycling Shorts	699.0	267.0
4	Sicons All Purpose Arnica Dog Shampoo	208	258	Sicons All Purpose Arnica Dog Shampoo	220.0	210.0

Method-2

Using merge to join columns

First combining the data using pandas merge features and removing the null values.

In [23]:
```
az_data=pd.read_csv('amz_com-ecommerce_sample.csv',encoding='unicode_escape')
fp_data=pd.read_csv('flipkart_com-ecommerce_sample.csv',encoding='unicode_escape')
```

In [24]:
```
az_data = az_data[["uniq_id","product_name","retail_price", "discounted_price"]]
az_data.rename(columns = {'product_name':'amazon_product_name', 'retail_price':'amazon_retail_price',
                          'discounted_price':'amazon_discounted_price'}, inplace = True)
az_data.head()
```

Out[24]:

	uniq_id	amazon_product_name	amazon_retail_price	amazon_discounted_price
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	982	438
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32143	29121
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	991	551
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	694	325
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	208	258

In [25]:

```python
fp_data = fp_data[["uniq_id","product_name","retail_price", "discounted_price"]]
fp_data.rename(columns = {'product_name':'flipkart_product_name', 'retail_price':'flipkart_retail_price',
                          'discounted_price':'flipkart_discounted_price'}, inplace = True)
fp_data.head()
```

Out[25]:

	uniq_id	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	999.0	379.0
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32157.0	22646.0
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	999.0	499.0
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	699.0	267.0
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	220.0	210.0

In [26]:

```python
new_data = az_data.merge(fp_data)
new_data.head()
```

Out[26]:

	uniq_id	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_
0	c2d766ca982eca8304150849735ffef9	Alisha Solid Women's Cycling Shorts	982	438	Alisha Solid Women's Cycling Shorts	999.0	
1	7f7036a6d550aaa89d34c77bd39a5e48	FabHomeDecor Fabric Double Sofa Bed	32143	29121	FabHomeDecor Fabric Double Sofa Bed	32157.0	
2	f449ec65dcbc041b6ae5e6a32717d01b	AW Bellies	991	551	AW Bellies	999.0	
3	0973b37acd0c664e3de26e97e5571454	Alisha Solid Women's Cycling Shorts	694	325	Alisha Solid Women's Cycling Shorts	699.0	
4	bc940ea42ee6bef5ac7cea3fb5cfbee7	Sicons All Purpose Arnica Dog Shampoo	208	258	Sicons All Purpose Arnica Dog Shampoo	220.0	

In [27]:

```python
new_data = new_data.drop(['uniq_id'],axis = 1)
new_data.head()
```

Out[27]:

	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
0	Alisha Solid Women's Cycling Shorts	982	438	Alisha Solid Women's Cycling Shorts	999.0	379.0
1	FabHomeDecor Fabric Double Sofa Bed	32143	29121	FabHomeDecor Fabric Double Sofa Bed	32157.0	22646.0
2	AW Bellies	991	551	AW Bellies	999.0	499.0
3	Alisha Solid Women's Cycling Shorts	694	325	Alisha Solid Women's Cycling Shorts	699.0	267.0
4	Sicons All Purpose Arnica Dog Shampoo	208	258	Sicons All Purpose Arnica Dog Shampoo	220.0	210.0

In [28]:

```python
new_data.shape
```

Out[28]:

```
(20000, 6)
```

In [29]:

```python
new_data.isnull().sum()
```

Out[29]:

```
amazon_product_name           0
amazon_retail_price           0
amazon_discounted_price       0
flipkart_product_name         0
flipkart_retail_price        78
flipkart_discounted_price    78
dtype: int64
```

In [30]:

```python
new_data=new_data.dropna()
new_data.head()
```

Out[30]:

	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
0	Alisha Solid Women's Cycling Shorts	982	438	Alisha Solid Women's Cycling Shorts	999.0	379.0
1	FabHomeDecor Fabric Double Sofa Bed	32143	29121	FabHomeDecor Fabric Double Sofa Bed	32157.0	22646.0
2	AW Bellies	991	551	AW Bellies	999.0	499.0
3	Alisha Solid Women's Cycling Shorts	694	325	Alisha Solid Women's Cycling Shorts	699.0	267.0
4	Sicons All Purpose Arnica Dog Shampoo	208	258	Sicons All Purpose Arnica Dog Shampoo	220.0	210.0

In [31]:

```
new_data.shape
```

Out[31]:

```
(19922, 6)
```

Shape or the size of data for the both the method are equal.

Building a user bot that takes product name as input and gives the required output

In [32]:

```
input1 = input("Enter product name: ")
print(input1)
```

```
Enter product name: AW Bellies
AW Bellies
```

In [33]:

```
Find_Data = new_data[new_data.isin([input1]).any(axis=1)]
Find_Data
```

Out[33]:

	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
2	AW Bellies	991	551	AW Bellies	999.0	499.0

In [34]:

```
input2 = input("Enter product name: ")
print(input2)
```

```
Enter product name: FDT Women's Leggings
FDT Women's Leggings
```

In [35]:

```
Find_Data = new_data[new_data.isin([input2]).any(axis=1)]
Find_Data
```

Out[35]:

	amazon_product_name	amazon_retail_price	amazon_discounted_price	flipkart_product_name	flipkart_retail_price	flipkart_discounted_price
28	FDT WOMEN'S Leggings Pants	698	362	FDT Women's Leggings	699.0	309.0