Fuzzy Cleaning

In [ ]:
```
import pandas as pd
Brand = pd.read_csv('../6Wresearch Data Analyst/Fuzzy Cleaning/bran
Brand.head()
```

Out[2]:

	slno	Brand Name
0	0	SONY
1	1	SONY
2	2	FUJIFILM
3	3	FUJIFILM
4	4	FUJIFILM

In [ ]:
```
Bank = pd.read_excel('../6Wresearch Data Analyst/Fuzzy Cleaning/dat
Bank.head()
```

Out[3]:

	Product Description	Model Name	Brand Name
0	1ST LENS ASSY (9147),A-2182-241-A,SEL1635GM(PA...	SEL1635GM	Sony
1	92992523 VPLL-Z4015/4WW SHORT FOCUS ZOOM LENS ...	VPLL-Z4015/4WW	Sony
2	16536659 FUJINON LENS GF32-64MMF4 R LM WR	GF32-64 mm	Fujifilm
3	16443060 F XF50-140MMF2.8 R LM OIS WR (LENS)	XF50-140 mm	Fujifilm
4	16536659 FUJINON LENS GF32-64MMF4 R LM WR	GF32-64 mm	Fujifilm

In [ ]:
```
Import = pd.read_excel('../6Wresearch Data Analyst/Fuzzy Cleaning/i
Import.head()
```

Out[4]:

	slno	Date	HS Code	Product Description	Consignee Name	Shipper Name	Consignee Address 1	Consign Address
0	0	2023-06-01	85258900	M320 IVCIII CAMERA - 4K 10449087 SPARE PARTS ...	DHR HOLDING INDIA PRIVATE LIMITED	LEICA MIKROSYSTEME VERTRIEB GMBH -DSA	UNIT NO 325 TO 328 DLF TOWERS,SHIV AJI MARG	Na
1	1	2023-06-01	85258900	16759732 FUJIFILM DIGITAL CAMERA F X-T30 II B/...	FUJIFILM INDIA PRIVATE LIMITED	FUJIFILM CORPORATION	BUSINESS CENTRE OFF NO 521 LEVEL V ,CADDIE COM...	Na
2	2	2023-06-01	85258900	16670041 FUJIFILM DIGITAL CAMERA F X-S10 CD LI...	FUJIFILM INDIA PRIVATE LIMITED	FUJIFILM CORPORATION	BUSINESS CENTRE OFF NO 521 LEVEL V ,CADDIE COM...	Na
3	3	2023-06-01	85258900	M320 IVCIII CAMERA - 4K 10449087 SPARE PARTS ...	DHR HOLDING INDIA PRIVATE LIMITED	LEICA MIKROSYSTEME VERTRIEB GMBH -DSA	UNIT NO 325 TO 328 DLF TOWERS,SHIV AJI MARG	Na
4	4	2023-06-01	85258900	16670041 FUJIFILM DIGITAL CAMERA F X-S10 CD LI...	FUJIFILM INDIA PRIVATE LIMITED	FUJIFILM CORPORATION	BUSINESS CENTRE OFF NO 521 LEVEL V ,CADDIE COM...	Na

5 rows × 52 columns

Data Cleaning

In [ ]:
```
Brand = Brand.drop(['slno'], axis=1)
Brand.head()
```

Out[5]:

	Brand Name
0	SONY
1	SONY
2	FUJIFILM
3	FUJIFILM
4	FUJIFILM

In [ ]:
```
Import = Import.drop(['slno', 'Date', 'HS Code', 'Consignee Name',
Import.head()
```

Out[6]:

	Product Description	Shipper Name	Shipper Address1	Shipper Address 2	Standard Qty	Standard Unit	Standard Unit Rate $	
0	M320 IVCIII CAMERA - 4K 10449087 SPARE PARTS ...	LEICA MIKROSYSTEME VERTRIEB GMBH -DSA	ERNST-LEITZ-STRASSE 17-3735578 WETZLAR / DUETS...	NaN	1	NOS	2101.580	3(
1	16759732 FUJIFILM DIGITAL CAMERA F X-T30 II B/...	FUJIFILM CORPORATION	26-30 NASHIAZABU,2 CHOME MINATO - KUTOKYO JAPAN...	NaN	2	NOS	505.000	1(
2	16670041 FUJIFILM DIGITAL CAMERA F X-S10 CD LI...	FUJIFILM CORPORATION	26-30 NASHIAZABU,2 CHOME MINATO - KUTOKYO JAPAN...	NaN	12	NOS	505.000	1
3	M320 IVCIII CAMERA - 4K 10449087 SPARE PARTS ...	LEICA MIKROSYSTEME VERTRIEB GMBH -DSA	ERNST-LEITZ-STRASSE 17-3735578 WETZLAR / DUETS...	NaN	1	NOS	2101.580	3(
4	16670041 FUJIFILM DIGITAL CAMERA F X-S10 CD LI...	FUJIFILM CORPORATION	26-30, NISHIAZABU 2-CHOME MINATO-KUTOKYO 106-8...	NaN	12	NOS	501.147	1

5 rows × 46 columns

Task 1 Return Brand Name and Model Name in import table from Data bank table where product description of import file is 100% match with product description of Data Bank.

In [ ]:
```
Hund = pd.merge(Import, Bank, how="inner", on="Product Description"
Hund[['Brand Name', 'Model Name']]
```

Out[7]:

	Brand Name	Model Name
0	Fujifilm	XF16 mm
1	Fujifilm	XF16 mm
2	Fujifilm	XF30 mm
3	Fujifilm	XF30 mm
4	Fujifilm	XF30 mm
...	...	...
2796	Sony	ILCE-7M3
2797	Sony	ZV-E10
2798	Sony	SRG-X120/BC
2799	Sony	SRG-X120/BC
2800	Sony	HDC3500/L4 UCJ

2801 rows × 2 columns

Answer of Task1

Task 3 – For rest rows of import table where Brand Name and Model name column is not returned from task1 and task 2, find Brand Name from 'Shipper Name' column and find it in Brand table.

In [ ]:
```
Import1 = Import[~Import['Product Description'].isin(Hund['Product 
Bank1 = Bank[~Bank['Product Description'].isin(Hund['Product Descri
```

In [ ]:
```
!pip install rapidfuzz
```

```
Requirement already satisfied: rapidfuzz in /usr/local/lib/python3
.10/dist-packages (3.1.2)
```

In [10]:
```
t = set()
s = set()
Nin_I = pd.DataFrame()
Nin_B = pd.DataFrame()
from rapidfuzz import fuzz
for i in Import1['Product Description'].index.values:
    for j in Bank1['Product Description'].index.values:
        if fuzz.ratio(Import1['Product Description'][i], Bank1['Pro
            try:
                Nin_I = Nin_I.append(Import1.iloc[i])
                Nin_B = Nin_B.append(Bank1.iloc[j])
            except:
                Nin_I = Nin_I.append(Import1.iloc[i])
                Nin_B = Nin_B.append(Bank1.loc[j])
```

```
Streaming output truncated to the last 5000 lines.
<ipython-input-10-170e680335fa>:8: FutureWarning: The frame.append
method is deprecated and will be removed from pandas in a future v
ersion. Use pandas.concat instead.
  Ten_I = Ten_I.append(Import1.iloc[i])
<ipython-input-10-170e680335fa>:9: FutureWarning: The frame.append
method is deprecated and will be removed from pandas in a future v
ersion. Use pandas.concat instead.
  Ten_B = Ten_B.append(Bank1.iloc[j])
<ipython-input-10-170e680335fa>:8: FutureWarning: The frame.append
method is deprecated and will be removed from pandas in a future v
ersion. Use pandas.concat instead.
  Ten_I = Ten_I.append(Import1.iloc[i])
<ipython-input-10-170e680335fa>:9: FutureWarning: The frame.append
method is deprecated and will be removed from pandas in a future v
ersion. Use pandas.concat instead.
  Ten_B = Ten_B.append(Bank1.iloc[j])
<ipython-input-10-170e680335fa>:8: FutureWarning: The frame.append
method is deprecated and will be removed from pandas in a future v
```

In [19]:
```
Import.columns
```

Out[19]:
```
Index(['Product Description', 'Shipper Name', 'Shipper Address1 ',
       'Shipper Address 2 ', 'Standard Qty', 'Standard Unit',
       'Standard Unit Rate $', 'Tax %', 'Estimated CIF Value $', '
Unit Rate $',
       'Port of Destination', 'Country of Origin', 'QTY', 'Unit',
'Rate In FC',
       'Rate Currency', 'Tax $', 'Landed Value $', 'Month', 'HS2',
'HS4',
       'Consignee City', 'Consignee Pincode', 'Consignee State',
       'Consignee Phone', 'Consignee E-mail', 'Contact Person', 'S
hipper City',
       'Shipper Country', 'Notify Party', 'BL TYP', 'Shipment Mode
',
       'Port Of Origin', 'HS Description', 'Raw Consignee Name',
       'Raw Shipper Name', 'Raw Consignee Add1', 'Raw Consignee Ad
d2',
       'Raw Shipper Address1', 'Raw Shipper Address2', 'Raw Consig
nee City',
       'Raw Consignee Pincode', 'Raw Consignee State', 'Raw Consig
nee Phone',
       'Raw Consignee E-mail', 'Record Id'],
      dtype='object')
```

In [20]:
```
Brand
```

Out[20]:

	Brand Name
0	SONY
1	SONY
2	FUJIFILM
3	FUJIFILM
4	FUJIFILM
...	...
20229	SONY
20230	FUJIFILM
20231	SONY
20232	SONY
20233	PANASONIC

20234 rows × 1 columns

In [24]:
```
Ten_I.columns
```

Out[24]:
```
Index(['Product Description', 'Shipper Name', 'Shipper Address1 ',
       'Shipper Address 2 ', 'Standard Qty', 'Standard Unit',
       'Standard Unit Rate $', 'Tax %', 'Estimated CIF Value $', '
Unit Rate $',
       'Port of Destination', 'Country of Origin', 'QTY', 'Unit',
'Rate In FC',
       'Rate Currency', 'Tax $', 'Landed Value $', 'Month', 'HS2',
'HS4',
       'Consignee City', 'Consignee Pincode', 'Consignee State',
       'Consignee Phone', 'Consignee E-mail', 'Contact Person', 'S
hipper City',
       'Shipper Country', 'Notify Party', 'BL TYP', 'Shipment Mode
',
       'Port Of Origin', 'HS Description', 'Raw Consignee Name',
       'Raw Shipper Name', 'Raw Consignee Add1', 'Raw Consignee Ad
d2',
       'Raw Shipper Address1', 'Raw Shipper Address2', 'Raw Consig
nee City',
       'Raw Consignee Pincode', 'Raw Consignee State', 'Raw Consig
nee Phone',
       'Raw Consignee E-mail', 'Record Id'],
      dtype='object')
```

In [29]:
```
Ten_I['Shipper Name']
```

Out[29]:
```
0       LEICA MIKROSYSTEME VERTRIEB GMBH -DSA
0       LEICA MIKROSYSTEME VERTRIEB GMBH -DSA
0       LEICA MIKROSYSTEME VERTRIEB GMBH -DSA
0       LEICA MIKROSYSTEME VERTRIEB GMBH -DSA
0       LEICA MIKROSYSTEME VERTRIEB GMBH -DSA
                      ...
1517    SONY ELECTRONICS ASIA PACIFIC PTE LTD
1517    SONY ELECTRONICS ASIA PACIFIC PTE LTD
1517    SONY ELECTRONICS ASIA PACIFIC PTE LTD
1517    SONY ELECTRONICS ASIA PACIFIC PTE LTD
1517    SONY ELECTRONICS ASIA PACIFIC PTE LTD
Name: Shipper Name, Length: 38294, dtype: object
```

In [31]:
```
Ten_B['Brand Name']
```

Out[31]:
```
15570    Tamron
15572    Tamron
15574    Tamron
15575    Tamron
15578    Tamron
          ...
18695     Nikon
18721     Nikon
18868     Nikon
18870     Nikon
18942     Nikon
Name: Brand Name, Length: 38294, dtype: object
```

Task 2 - Return Brand Name and Model name column in import table from Data bank table where product description of import file is up to 90% match with product description of Data Bank. (Consider only those product descriptions of import table for which Brand Name and Model name column is not returned in task 1).

In [14]:
```
Import_Nin = Import1[~Import1['Product Description'].isin(Ten_I['Pr
Bank_Nin = Bank1[~Bank1['Product Description'].isin(Ten_B['Product
```

In [17]:
```
Bank_Nin[['Brand Name', 'Model Name']]
```

Out[17]:

	Brand Name	Model Name
1	Sony	VPLL-Z4015/4WW
2	Fujifilm	GF32-64 mm
4	Fujifilm	GF32-64 mm
6	Nikon	MQA18000 DS-FI3
7	Fujifilm	F1B013039220
...	...	...
20230	Sony	F XT200
20231	Fujifilm	X100
20232	Sony	FX-X-T3
20233	Sony	F FX-X
20234	Panasonic	PI-SFW103L

5626 rows × 2 columns