Nayan Ranjan Das | A2305220148 | 5CSE2

Lab 1

August 8, 2022

---

importing required Python libraries

In [1]:

```
import numpy as np
import pandas as pd
```

reading the datasets

In [2]:

```
cust_df   = pd.read_csv('global_sales_data/cust_dimen.csv')
market_df = pd.read_csv('global_sales_data/market_fact.csv')
orders_df = pd.read_csv('global_sales_data/orders_dimen.csv')
prod_df   = pd.read_csv('global_sales_data/prod_dimen.csv')
shoip_df  = pd.read_csv('global_sales_data/shipping_dimen.csv')
```

In [3]:

```
cust_df.head()
```

Out[3]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
0	MUHAMMED MACINTYRE	NUNAVUT	NUNAVUT	SMALL BUSINESS	Cust_1
1	BARRY FRENCH	NUNAVUT	NUNAVUT	CONSUMER	Cust_2
2	CLAY ROZENDAL	NUNAVUT	NUNAVUT	CORPORATE	Cust_3
3	CARLOS SOLTERO	NUNAVUT	NUNAVUT	CONSUMER	Cust_4
4	CARL JACKSON	NUNAVUT	NUNAVUT	CORPORATE	Cust_5

selection of rows and columns using `loc` and `iloc`

In [9]:

```
1 cust_df.iloc[3,0]
```

Out[9]:

```
'CARLOS SOLTERO'
```

In [8]:

```
1 cust_df.loc[3,'Customer_Name']
```

Out[8]:

```
'CARLOS SOLTERO'
```

In [12]:

```
1 cust_df.iloc[[2,3],:]
```

Out[12]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
2	CLAY ROZENDAL	NUNAVUT	NUNAVUT	CORPORATE	Cust_3
3	CARLOS SOLTERO	NUNAVUT	NUNAVUT	CONSUMER	Cust_4

In [13]:

```
1 cust_df.iloc[:5,:]
```

Out[13]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
0	MUHAMMED MACINTYRE	NUNAVUT	NUNAVUT	SMALL BUSINESS	Cust_1
1	BARRY FRENCH	NUNAVUT	NUNAVUT	CONSUMER	Cust_2
2	CLAY ROZENDAL	NUNAVUT	NUNAVUT	CORPORATE	Cust_3
3	CARLOS SOLTERO	NUNAVUT	NUNAVUT	CONSUMER	Cust_4
4	CARL JACKSON	NUNAVUT	NUNAVUT	CORPORATE	Cust_5

note:

- `loc` gives values including the last index
- `iloc` doesn't

In [14]:

```
1 cust_df.loc[:5,:]
```

Out[14]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
0	MUHAMMED MACINTYRE	NUNAVUT	NUNAVUT	SMALL BUSINESS	Cust_1
1	BARRY FRENCH	NUNAVUT	NUNAVUT	CONSUMER	Cust_2
2	CLAY ROZENDAL	NUNAVUT	NUNAVUT	CORPORATE	Cust_3
3	CARLOS SOLTERO	NUNAVUT	NUNAVUT	CONSUMER	Cust_4
4	CARL JACKSON	NUNAVUT	NUNAVUT	CORPORATE	Cust_5
5	MONICA FEDERLE	NUNAVUT	NUNAVUT	CORPORATE	Cust_6

In [17]:

```
1 cust_df.loc[5:8]
```

Out[17]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
5	MONICA FEDERLE	NUNAVUT	NUNAVUT	CORPORATE	Cust_6
6	DOROTHY BADDERS	NUNAVUT	NUNAVUT	HOME OFFICE	Cust_7
7	NEOLA SCHNEIDER	NUNAVUT	NUNAVUT	HOME OFFICE	Cust_8
8	CARLOS DALY	NUNAVUT	NUNAVUT	HOME OFFICE	Cust_9

In [18]:

```
1 cust_df.iloc[[3,7,8]]
```

Out[18]:

	Customer_Name	Province	Region	Customer_Segment	Cust_id
3	CARLOS SOLTERO	NUNAVUT	NUNAVUT	CONSUMER	Cust_4
7	NEOLA SCHNEIDER	NUNAVUT	NUNAVUT	HOME OFFICE	Cust_8
8	CARLOS DALY	NUNAVUT	NUNAVUT	HOME OFFICE	Cust_9

In [19]:

```
1 market_df.iloc[3,2]
```

Out[19]:

```
'SHP_7625'
```

In [21]:

```
1 market_df.iloc[[3,5,6]]
```

Out[21]:

	Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit	Shippi
3	Ord_5456	Prod_6	SHP_7625	Cust_1818	2337.89	0.09	43	729.34	
5	Ord_5446	Prod_6	SHP_7608	Cust_1818	164.02	0.03	23	-47.64	
6	Ord_31	Prod_12	SHP_41	Cust_26	14.76	0.01	5	1.32	

In [4]:

```
1 market_df.loc[[3*2]]
```

Out[4]:

	Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit	Shipping_Cost
6	Ord_31	Prod_12	SHP_41	Cust_26	14.76	0.01	5	1.32	0.5

changing index of dataframe

In [40]:

```
1 market_order = market_df.set_index('Ord_id')
```

In [46]:

```
1 market_order.loc[['Ord_5406','Ord_5485'],]
```

Out[46]:

Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit	Shipping
Ord_5406	Prod_13	SHP_7549	Cust_1818	42.27	0.01	13	4.56	
Ord_5485	Prod_17	SHP_7664	Cust_1818	4233.15	0.08	35	1219.87	

In [56]:

```
1 market_df['Sales']
```

Out[56]:

```
0        136.8100
1         42.2700
2       4701.6900
3       2337.8900
4       4233.1500
          ...
8394    2841.4395
8395     127.1600
8396     243.0500
8397    3872.8700
8398     603.6900
Name: Sales, Length: 8399, dtype: float64
```

Slicing and dicing dataframe

In [50]:

```
1 market_df['Sales'] > 2000
```

Out[50]:

```
0       False
1       False
2        True
3        True
4        True
        ...
8394     True
8395    False
8396    False
8397     True
8398    False
Name: Sales, Length: 8399, dtype: bool
```

In [64]:

```
1 market_df.Discount.describe()
```

Out[64]:

```
count    8399.000000
mean        0.049671
std         0.031823
min         0.000000
25%         0.020000
50%         0.050000
75%         0.080000
max         0.250000
Name: Discount, dtype: float64
```

In [61]:

```
1 market_df.loc[market_df.Sales > 2000]
```

Out[61]:

	Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit
2	Ord_5446	Prod_4	SHP_7610	Cust_1818	4701.6900	0.00	26	1148.90
3	Ord_5456	Prod_6	SHP_7625	Cust_1818	2337.8900	0.09	43	729.34
4	Ord_5485	Prod_17	SHP_7664	Cust_1818	4233.1500	0.08	35	1219.87
7	Ord_4725	Prod_4	SHP_6593	Cust_1641	3410.1575	0.10	48	1137.91
10	Ord_4743	Prod_2	SHP_6615	Cust_1641	4072.0100	0.01	43	1675.98
...	...	...	...	...	...	...	...	...
8371	Ord_2624	Prod_4	SHP_3591	Cust_1006	4924.1350	0.07	28	1049.54
8381	Ord_2696	Prod_4	SHP_3691	Cust_1006	2836.0505	0.01	25	561.13
8383	Ord_2722	Prod_1	SHP_3731	Cust_1006	3508.3300	0.04	21	-546.98
8394	Ord_5353	Prod_4	SHP_7479	Cust_1798	2841.4395	0.08	28	374.63
8397	Ord_5348	Prod_15	SHP_7469	Cust_1798	3872.8700	0.03	23	565.34

1872 rows × 10 columns

In [73]:

```
1 market_df.loc[(market_df.Sales > 2000) & (market_df.Sales < 3000) & (market_df.Profit
```

Out[73]:

	Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit	S
3	Ord_5456	Prod_6	SHP_7625	Cust_1818	2337.8900	0.09	43	729.34	
81	Ord_5205	Prod_4	SHP_7274	Cust_1749	2546.5235	0.09	26	210.00	
109	Ord_139	Prod_17	SHP_186	Cust_45	2671.2100	0.06	14	636.18	
110	Ord_239	Prod_4	SHP_332	Cust_45	2157.3085	0.00	38	519.25	
141	Ord_1673	Prod_17	SHP_2314	Cust_498	2027.5500	0.04	14	537.40	
...	...	...	...	...	...	...	...	...	
8338	Ord_2107	Prod_2	SHP_2882	Cust_785	2409.9600	0.07	32	575.10	
8350	Ord_3570	Prod_4	SHP_4942	Cust_1266	2094.9780	0.06	44	697.29	
8354	Ord_3592	Prod_4	SHP_4973	Cust_1266	2614.3705	0.07	25	384.01	
8381	Ord_2696	Prod_4	SHP_3691	Cust_1006	2836.0505	0.01	25	561.13	
8394	Ord_5353	Prod_4	SHP_7479	Cust_1798	2841.4395	0.08	28	374.63	

328 rows × 10 columns

In [74]:

```
1 market_df.loc[(market_df.Discount > 0.15) | market_df.Sales > 2000]
```

Out[74]:

Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit	Shipping_Cost	P

In [76]:

```
1 market_df.loc[~(market_df.Sales < 3000)]
```

Out[76]:

	Ord_id	Prod_id	Ship_id	Cust_id	Sales	Discount	Order_Quantity	Profit
2	Ord_5446	Prod_4	SHP_7610	Cust_1818	4701.6900	0.00	26	1148.90
4	Ord_5485	Prod_17	SHP_7664	Cust_1818	4233.1500	0.08	35	1219.87
7	Ord_4725	Prod_4	SHP_6593	Cust_1641	3410.1575	0.10	48	1137.91
10	Ord_4743	Prod_2	SHP_6615	Cust_1641	4072.0100	0.01	43	1675.98
13	Ord_2207	Prod_11	SHP_3093	Cust_839	3364.2480	0.10	15	-693.23
...	...	...	...	...	...	...	...	...
8366	Ord_3593	Prod_3	SHP_4974	Cust_1274	12073.0600	0.03	39	5081.87
8367	Ord_3593	Prod_15	SHP_4975	Cust_1274	6685.0500	0.09	25	1653.60
8371	Ord_2624	Prod_4	SHP_3591	Cust_1006	4924.1350	0.07	28	1049.54
8383	Ord_2722	Prod_1	SHP_3731	Cust_1006	3508.3300	0.04	21	-546.98
8397	Ord_5348	Prod_15	SHP_7469	Cust_1798	3872.8700	0.03	23	565.34

1359 rows × 10 columns