Record Matching

> There are 2 datasets present in the file. Data 1 and Data 2 The primary key for both data1 and data2 is Order Id + Product ID combination (i.e. the individual datasets do not have any duplicate on this combination)

1. How to identify the Records (Order ID + Product ID combination) present in data1 but missing in data2 (Specify the number of records missing in your answer)

2. How to identify the Records (Order ID + Product ID combination) missing in data1 but present in data2 (Specify the number of records missing in your answer)

3. Find the Sum of the total Qty of Records missing in data1 but present in data2

4. Find the total number of unique records (Order ID + Product ID combination) present in the combined dataset of data1 and data2

In [3]:

```
import pandas as pd
#df = pd.read_excel(r"C:\Users\PC-chetan\Downloads\Records Matching Task.xlsx")
df1 = pd.read_csv(r"C:\Users\PC-chetan\Downloads\Records Matching Task.xlsx - data1.csv")
df2 = pd.read_csv(r"C:\Users\PC-chetan\Downloads\Records Matching Task.xlsx - data2.csv")
```

> Records (Order ID + Product ID combination) present in data1 but missing in data2

In [32]:

```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9478 entries, 0 to 9477
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Order ID    9478 non-null   object
 1   Product ID  9478 non-null   object
 2   Qty         9478 non-null   int64 
dtypes: int64(1), object(2)
memory usage: 222.3+ KB
```

In [33]:

```
df2.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9479 entries, 0 to 9478
Data columns (total 3 columns):
 #   Column      Non-Null Count  Dtype 
---  ------      --------------  ----- 
 0   Order ID    9479 non-null   object
 1   Product ID  9479 non-null   object
 2   Qty         9479 non-null   int64 
dtypes: int64(1), object(2)
memory usage: 222.3+ KB
```

In [37]:

```
df1.apply(tuple, 1)
```

Loading [MathJax]/extensions/Safe.js

Out[37]:
```
0       (CA-2014-100006, TEC-PH-10002075, 3)
1       (CA-2014-100090, FUR-TA-10003715, 3)
2       (CA-2014-100090, OFF-BI-10001597, 6)
3       (CA-2014-100293, OFF-PA-10000176, 6)
4       (CA-2014-100328, OFF-BI-10000343, 1)
                         ...
9473    (US-2017-169551, OFF-PA-10004100, 3)
9474    (US-2017-169551, OFF-ST-10004835, 3)
9475    (US-2017-169551, TEC-AC-10002018, 3)
9476    (US-2017-169551, TEC-AC-10003033, 2)
9477    (US-2017-169551, TEC-PH-10001363, 2)
Length: 9478, dtype: object
```

In [7]:
```
result1 = df1[~df1.apply(tuple, 1).isin(df2.apply(tuple, 1))]
```

In [8]:
```
result1
```

Out[8]:

	Order ID	Product ID	Qty
0	CA-2014-100006	TEC-PH-10002075	3
10	CA-2014-100678	OFF-EN-10000056	3
19	CA-2014-100895	OFF-AR-10004511	2
35	CA-2014-101560	OFF-BI-10000309	3
61	CA-2014-102673	OFF-LA-10001771	12
...	...	...	...
9390	US-2017-160836	OFF-AP-10001626	2
9403	US-2017-162558	FUR-FU-10002364	2
9420	US-2017-163657	OFF-BI-10000138	5
9427	US-2017-164056	FUR-TA-10001307	5
9435	US-2017-165456	FUR-CH-10003981	6

507 rows × 3 columns

In [41]:
```
result1.count()
```

Out[41]:
```
Order ID      507
Product ID    507
Qty           507
dtype: int64
```

> Records (Order ID + Product ID combination) missing in data1 but present in data2

In [10]:
```
result2 = df2[~df2.apply(tuple, 1).isin(df1.apply(tuple, 1))]
```

In [11]:
```
result2
```

Out[11]:

	Order ID	Product ID	Qty
	706	TEC-AC-10001314	2

Loading [MathJax]/extensions/Safe.js

	Order ID	Product ID	Qty
14	CA-2014-100762	OFF-PA-10001815	3
30	CA-2014-101427	OFF-AR-10002257	3
56	CA-2014-102652	FUR-FU-10001918	7
63	CA-2014-102869	OFF-PA-10000788	3
...	...	...	...
9428	US-2017-165344	OFF-BI-10003196	10
9433	US-2017-165358	TEC-CO-10001943	5
9455	US-2017-167920	OFF-AP-10000159	5
9471	US-2017-169502	OFF-AP-10001947	5
9473	US-2017-169551	FUR-BO-10001519	3

508 rows × 3 columns

In [42]:

```
result2.count()
```

Out[42]:

```
Order ID      508
Product ID    508
Qty           508
dtype: int64
```

> Sum of the total Qty of Records missing in data1 but present in data2

In [23]:

```
result2.Qty.sum()
```

Out[23]:

```
1956
```

> Total number of unique records (Order ID + Product ID combination) present in the combined dataset of data1 and data2

In [52]:

```
u_record = pd.merge(result1,result2, how = "outer")
u_record
```

Out[52]:

	Order ID	Product ID	Qty
0	CA-2014-100006	TEC-PH-10002075	3
1	CA-2014-100678	OFF-EN-10000056	3
2	CA-2014-100895	OFF-AR-10004511	2
3	CA-2014-101560	OFF-BI-10000309	3
4	CA-2014-102673	OFF-LA-10001771	12
...	...	...	...
1010	US-2017-165344	OFF-BI-10003196	10

	Order ID	Product ID	Qty
1011	US-2017-165358	TEC-CO-10001943	5
1012	US-2017-167920	OFF-AP-10000159	5
1013	US-2017-169502	OFF-AP-10001947	5
1014	US-2017-169551	FUR-BO-10001519	3

1015 rows × 3 columns

In [53]:

```
u_record.count()
```

Out[53]:

```
Order ID      1015
Product ID    1015
Qty           1015
dtype: int64
```

Loading [MathJax]/extensions/Safe.js