In [2]:
```
import pandas as pd
```

In [3]:
```
trans_data =
pd.read_excel(r"C:\\Users\\LORD\\Downloads\\QVI_transaction_data.xlsx")
```

In [4]:
```
purchase_data =
pd.read_csv(r"C:\\Users\\LORD\\Downloads\\QVI_purchase_behaviour.csv")
```

In [15]:
```
trans_data
```

Out[15]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PRO
0	1970-01-01 00:00:00.000043390	1	1000	1	5	Natural Chip Compny SeaSalt175g	
1	1970-01-01 00:00:00.000043599	1	1307	348	66	CCs Nacho Cheese 175g	
2	1970-01-01 00:00:00.000043605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	
3	1970-01-01 00:00:00.000043329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	
4	1970-01-01 00:00:00.000043330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	
...	...	...	...	...	...	...	
264831	1970-01-01 00:00:00.000043533	272	272319	270088	89	Kettle Sweet Chilli And Sour Cream 175g	
264832	1970-01-01 00:00:00.000043325	272	272358	270154	74	Tostitos Splash Of Lime 175g	
264833	1970-01-01 00:00:00.000043410	272	272379	270187	51	Doritos Mexicana 170g	
264834	1970-01-01 00:00:00.000043461	272	272379	270188	42	Doritos Corn Chip Mexican Jalapeno 150g	
264835	1970-01-01 00:00:00.000043365	272	272380	270189	74	Tostitos Splash Of Lime 175g	

264836 rows × 8 columns

In [12]:
```
trans_data['DATE'] = pd.to_datetime(trans_data['DATE'])
trans_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264836 entries, 0 to 264835
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype         
---  ------          --------------   -----         
 0   DATE            264836 non-null  datetime64[ns]
 1   STORE_NBR       264836 non-null  int64         
 2   LYLTY_CARD_NBR  264836 non-null  int64         
 3   TXN_ID          264836 non-null  int64         
 4   PROD_NBR        264836 non-null  int64         
 5   PROD_NAME       264836 non-null  object        
 6   PROD_QTY        264836 non-null  int64         
 7   TOT_SALES       264836 non-null  float64       
dtypes: datetime64[ns](1), float64(1), int64(5), object(1)
memory usage: 16.2+ MB
```

In [29]:
```
trans_cleaned =
trans_data[trans_data['PROD_NAME'].str.contains(r'Chip|Chps', na = False)]
trans_cleaned['TOT_SALES'].max()
```

In [47]:
```
trans_cleaned[trans_cleaned['TOT_SALES'].isin(['NAN','',0])]
trans_cleaned[trans_cleaned['TOT_SALES'] <= 0]
```

Out[47]:

DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES

In [48]:
```
trans_cleaned.head(4)
```

Out[48]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY
0	1970-01-01 00:00:00.000043390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2
2	1970-01-01 00:00:00.000043605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2
3	1970-01-01 00:00:00.000043329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5
4	1970-01-01 00:00:00.000043330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3

In [54]:
```
purchase_data.drop_duplicates(subset ='LYLTY_CARD_NBR')
```

Out[54]:

	LYLTY_CARD_NBR	LIFESTAGE	PREMIUM_CUSTOMER
0	1000	YOUNG SINGLES/COUPLES	Premium
1	1002	YOUNG SINGLES/COUPLES	Mainstream
2	1003	YOUNG FAMILIES	Budget
3	1004	OLDER SINGLES/COUPLES	Mainstream
4	1005	MIDAGE SINGLES/COUPLES	Mainstream
...	...	...	...
72632	2370651	MIDAGE SINGLES/COUPLES	Mainstream
72633	2370701	YOUNG FAMILIES	Mainstream
72634	2370751	YOUNG FAMILIES	Premium
72635	2370961	OLDER FAMILIES	Budget
72636	2373711	YOUNG SINGLES/COUPLES	Mainstream

72637 rows × 3 columns

In [55]:
```
full_data = pd.merge(purchase_data , trans_cleaned , on =
'LYLTY_CARD_NBR')
```

In [97]:
```
Total_sales = round(full_data['TOT_SALES'].sum(),2)
slaes_per_store_nmbr = full_data.groupby('STORE_NBR',as_index = False)
['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' , ascending = False)
count_of_txnid_per_customer = full_data.groupby('LYLTY_CARD_NBR' ,
as_index = False)['TXN_ID'].count().sort_values(by = 'TXN_ID' , ascending
= False)
sum_of_sales_per_customer = full_data.groupby('LYLTY_CARD_NBR' , as_index
= False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' , ascending =
False)
sum_of_sales_per_life_stage = full_data.groupby('LIFESTAGE' , as_index =
False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' , ascending =
False)
sum_of_sales_per_premium_customer = full_data.groupby('PREMIUM_CUSTOMER' ,
as_index = False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' ,
ascending = False)
sum_of_sales_per_product = full_data.groupby('PROD_NAME' , as_index =
False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' , ascending =
False)
```

```
sum_of_sales_per_pack_size = full_data.groupby('pack_size' , as_index =
False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' , ascending =
False)
sum_of_sales_per_company_name = full_data.groupby('company_name' ,
as_index = False)['TOT_SALES'].sum().sort_values(by = 'TOT_SALES' ,
ascending = False)
```

In [94]:
```
full_data[['pack_name' ,'pack_size']] =
full_data['PROD_NAME'].str.split(r'(?=\d)' , n = 1 , expand = True)
full_data['company_name'] = full_data['pack_name'].str.split().str[0]
full_data.head(4)
```

Out[94]:

	LYLTY_CARD_NBR	LIFESTAGE	PREMIUM_CUSTOMER	DATE	STORE_NBR	TXN_ID
0	1000	YOUNG SINGLES/COUPLES	Premium	1970-01-01 00:00:00.000043390	1	
1	1003	YOUNG FAMILIES	Budget	1970-01-01 00:00:00.000043532	1	
2	1004	OLDER SINGLES/COUPLES	Mainstream	1970-01-01 00:00:00.000043406	1	
3	1011	OLDER SINGLES/COUPLES	Mainstream	1970-01-01 00:00:00.000043453	1	

In [99]:
```
full_data.to_csv(r'C:\\Users\\LORD\\Downloads\\job_sim2.csv')
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]: