In [21]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Transaction data set

In [22]:
```
TR = pd.read_excel('QVI_transaction_data.xlsx')
```

In [23]:
```
TR.head()
```

Out[23]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	T
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	

Summary

In [24]:
```
TR.describe()
```

Out[24]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	264836.000000	264836.00
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	56.583157	1.90
std	105.389282	76.78418	8.057998e+04	7.813303e+04	32.826638	0.64
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	1.000000	1.00
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	28.000000	2.00
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	56.000000	2.00
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	85.000000	2.00
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	114.000000	200.00

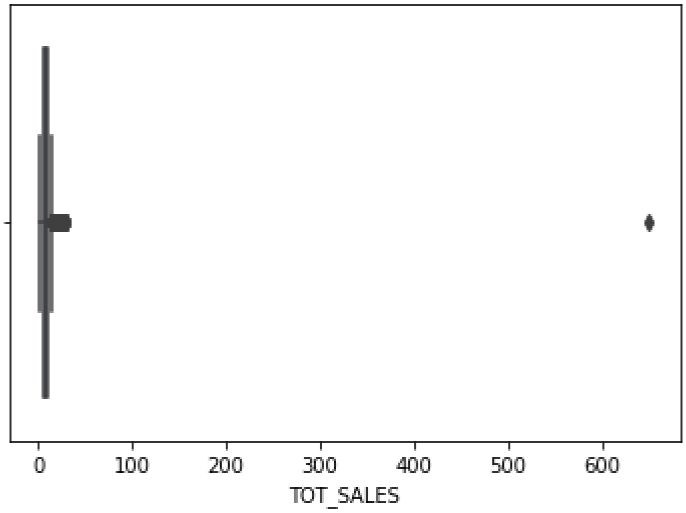
In [25]:
```
TR.isnull().sum()
```

Out[25]:
```
DATE              0
STORE_NBR         0
LYLTY_CARD_NBR    0
TXN_ID            0
PROD_NBR          0
PROD_NAME         0
PROD_QTY          0
TOT_SALES         0
dtype: int64
```

Identifying outliers

In [26]:
```
sns.boxplot(TR['TOT_SALES'])
```

Out[26]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x19f6d425248>
```



There is an outlier after the value of 600

Removing outliers

In [27]:
```
TR.sort_values(by='TOT_SALES', ascending = False)
```

Out[27]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY
69762	43331	226	226000	226201	4	Dorito Corn Chp Supreme 380g	200
69763	43605	226	226000	226210	4	Dorito Corn Chp Supreme 380g	200
69496	43327	49	49303	45789	14	Smiths Crnkle Chip Orgnl Big Bag 380g	5
55558	43599	190	190113	190914	14	Smiths Crnkle Chip Orgnl Big Bag 380g	5
171815	43329	24	24095	20797	14	Smiths Crnkle Chip Orgnl Big Bag 380g	5
...	...	...	...	...	...	...	...
259695	43417	41	41089	38002	76	Woolworths Medium Salsa 300g	1
259707	43391	41	41267	38201	76	Woolworths Medium Salsa 300g	1
197005	43323	167	167121	168928	76	Woolworths Medium Salsa 300g	1
216449	43525	264	264032	262778	76	Woolworths Medium Salsa 300g	1
150019	43405	268	268303	264733	35	Woolworths Mild Salsa 300g	1

264836 rows × 8 columns

In [29]:
```
a = TR[TR['TOT_SALES']>8.00].index
```

In [30]:
```
print(a)
TR.drop(a,inplace=True)
```

```
Int64Index([     3,      4,     11,     12,     16,     24,     31,     56,
                58,     65,
            ...
            264801, 264807, 264808, 264809, 264819, 264821, 264823, 264831,
            264833, 264835],
           dtype='int64', length=97934)
```

In [31]:
```
TR.sort_values(by='TOT_SALES', ascending = False)
```

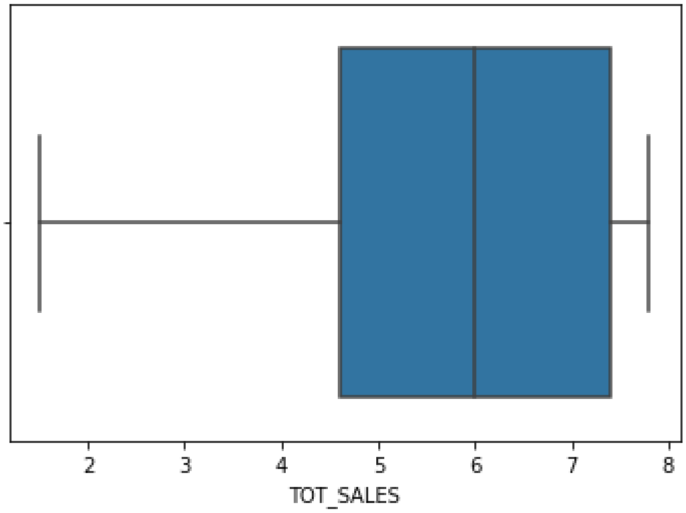
Out[31]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY
264834	43461	272	272379	270188	42	Doritos Corn Chip Mexican Jalapeno 150g	2
124675	43357	105	105162	106269	93	Doritos Corn Chip Southern Chicken 150g	2
67328	43633	226	226116	226823	42	Doritos Corn Chip Mexican Jalapeno 150g	2
251920	43284	180	180098	181613	93	Doritos Corn Chip Southern Chicken 150g	2
124765	43572	106	106090	107292	42	Doritos Corn Chip Mexican Jalapeno 150g	2
...	...	...	...	...	...	...	...
152264	43401	16	16287	14414	35	Woolworths Mild Salsa 300g	1
43380	43417	120	120140	123649	76	Woolworths Medium Salsa 300g	1
163352	43464	163	163153	163444	35	Woolworths Mild Salsa 300g	1
82497	43309	20	20416	17412	35	Woolworths Mild Salsa 300g	1
233083	43521	124	124184	127927	35	Woolworths Mild Salsa 300g	1

166902 rows × 8 columns

```
In [32]: sns.boxplot(TR['TOT_SALES'])

Out[32]: <matplotlib.axes._subplots.AxesSubplot at 0x19f76e2c2c8>
```



Now, there are no outliers

Data Formats

```
In [33]: TR.dtypes

Out[33]: DATE               int64
         STORE_NBR          int64
         LYLTY_CARD_NBR     int64
         TXN_ID             int64
         PROD_NBR           int64
         PROD_NAME         object
         PROD_QTY           int64
         TOT_SALES        float64
         dtype: object
```

Purchase Behavior Dataset

```
In [34]: PB = pd.read_csv('QVI_purchase_behaviour.csv')
In [35]: PB.head()
```

Out[35]:

	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

Summary

```
In [37]: PB.describe(include=object)
```

Out[37]:

	LIFESTAGE	PREMIUM_CUSTOMER
count	72637	72637
unique	7	3
top	RETIREES	Mainstream
freq	14805	29245

```
In [40]: PB.isnull().sum()
```

```
Out[40]: LYLTY_CARD_NBR      0
         LIFESTAGE           0
         PREMIUM_CUSTOMER    0
         dtype: int64
```

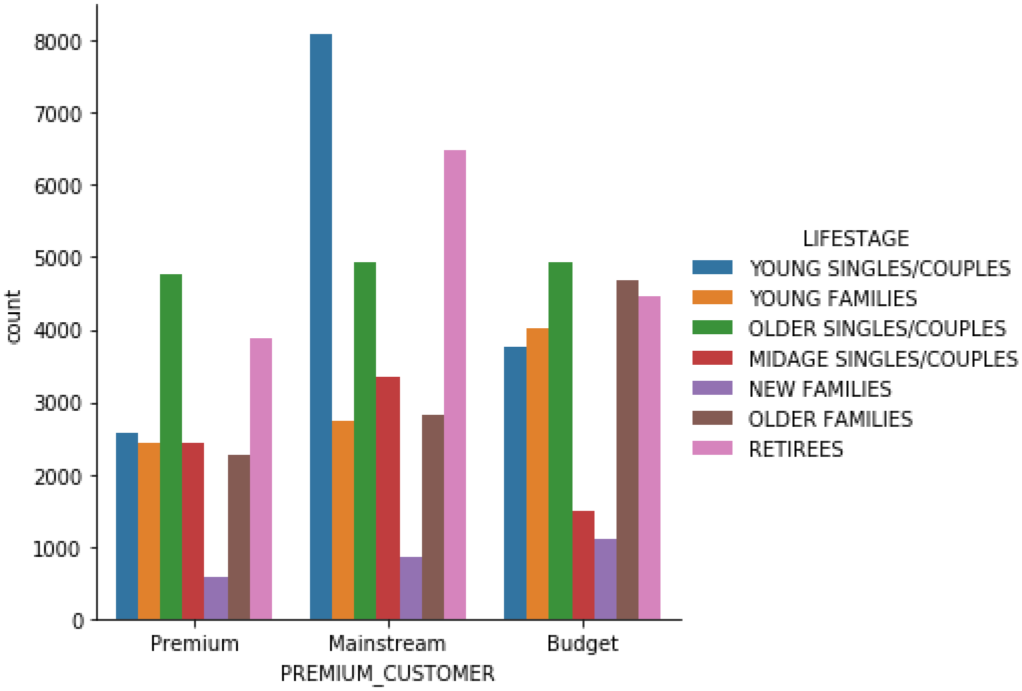
Outlier detection - found nothing

In [44]:

```
sns.catplot(x='PREMIUM_CUSTOMER', hue='LIFESTAGE', data = PB, kind = 'count')
```

Out[44]:

```
<seaborn.axisgrid.FacetGrid at 0x19f6e761348>
```



Data Format

In [45]:

```
PB.dtypes
```

Out[45]:

```
LYLTY_CARD_NBR       int64
LIFESTAGE           object
PREMIUM_CUSTOMER    object
dtype: object
```

In [ ]:

```

```