Quantium Data Analytics Virtual Internship

In [ ]:

```

```

Importing Libraries

In [4]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import plotly.express as pe
```

Importing Datasets

In [7]:

```
transaction_data = pd.read_excel("E:\Virtual Internship Data Analytics\Quantium Virtual I
purchase_behaviour = pd.read_csv("E:\Virtual Internship Data Analytics\Quantium Virtual I
```

In [9]:

```
transaction_data.head()
```

Out[9]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.0
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.8

In [10]:

```
purchase_behaviour.head()
```

Out[10]:

	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

In [ ]:

```

```

Creating Summaries of Datasets

In [11]:
```
transaction_data.describe()
```

Out[11]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_S
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	264836.000000	264836.000000	264836.0
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	56.583157	1.907309	7.3
std	105.389282	76.78418	8.057998e+04	7.813303e+04	32.826638	0.643654	3.0
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	1.000000	1.000000	1.5
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	28.000000	2.000000	5.4
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	56.000000	2.000000	7.4
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	85.000000	2.000000	9.2
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	114.000000	200.000000	650.0

In [12]:
```
purchase_behaviour.describe()
```

Out[12]:

	LYLTY_CARD_NBR
count	7.263700e+04
mean	1.361859e+05
std	8.989293e+04
min	1.000000e+03
25%	6.620200e+04
50%	1.340400e+05
75%	2.033750e+05
max	2.373711e+06

In [ ]:

Finding Null Values

In [14]:
```
transaction_data.isnull().sum()
```

Out[14]:
```
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
```

In [16]:
```
purchase_behaviour.isnull().sum()
```

Out[16]:
```
LYLTY_CARD_NBR      0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

In [ ]:

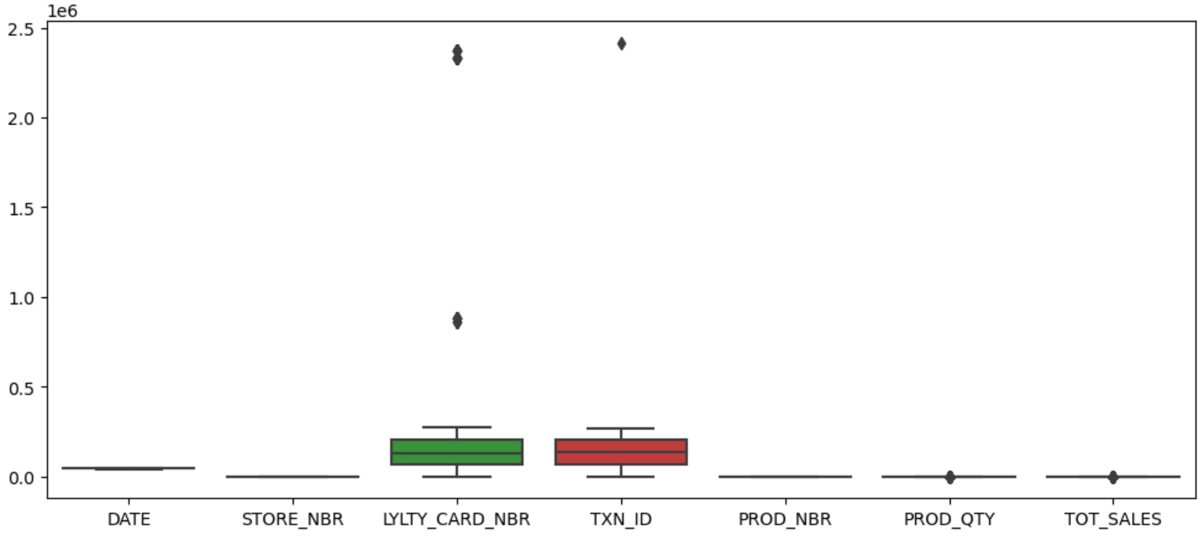
Finding Outliers

In [19]:
```
transaction_data.head()
```

Out[19]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALES
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.0
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.3
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.9
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.0
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.8

In [30]:
```
plt.figure(figsize = (12,5))
sns.boxplot(transaction_data);
```



In [32]:

```
transaction_data[['LYLTY_CARD_NBR']].sort_values('LYLTY_CARD_NBR',ascending =False).head(
```

Out[32]:

```
        LYLTY_CARD_NBR
256040         2373711
 53107         2370961
 53106         2370961
227371         2370751
215522         2370701
 15676         2370651
 97172         2370581
 97173         2370581
 97171         2370361
255925         2370181
133253         2370001
 96939         2330501
 32030         2330461
104927         2330431
135105         2330331
244444         2330321
228460         2330311
115267         2330291
 99033         2330291
 99034         2330291
```

In [25]:

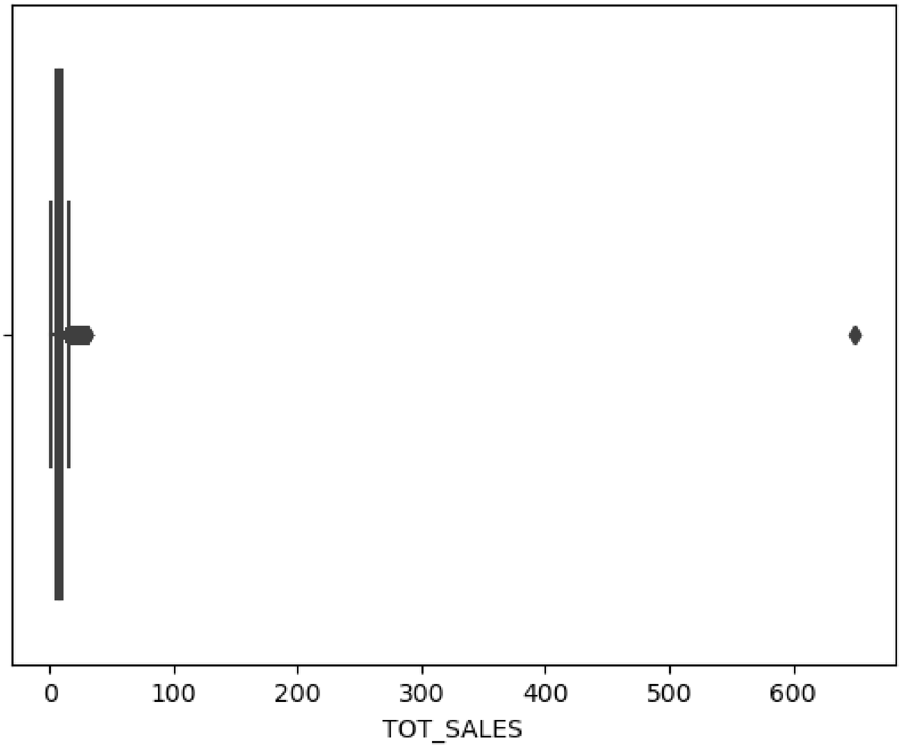
```
transaction_data.columns
```

Out[25]:

```
Index(['DATE', 'STORE_NBR', 'LYLTY_CARD_NBR', 'TXN_ID', 'PROD_NBR',
       'PROD_NAME', 'PROD_QTY', 'TOT_SALES'],
      dtype='object')
```

```
In [35]: sns.boxplot(x = 'TOT_SALES',data = transaction_data)
```

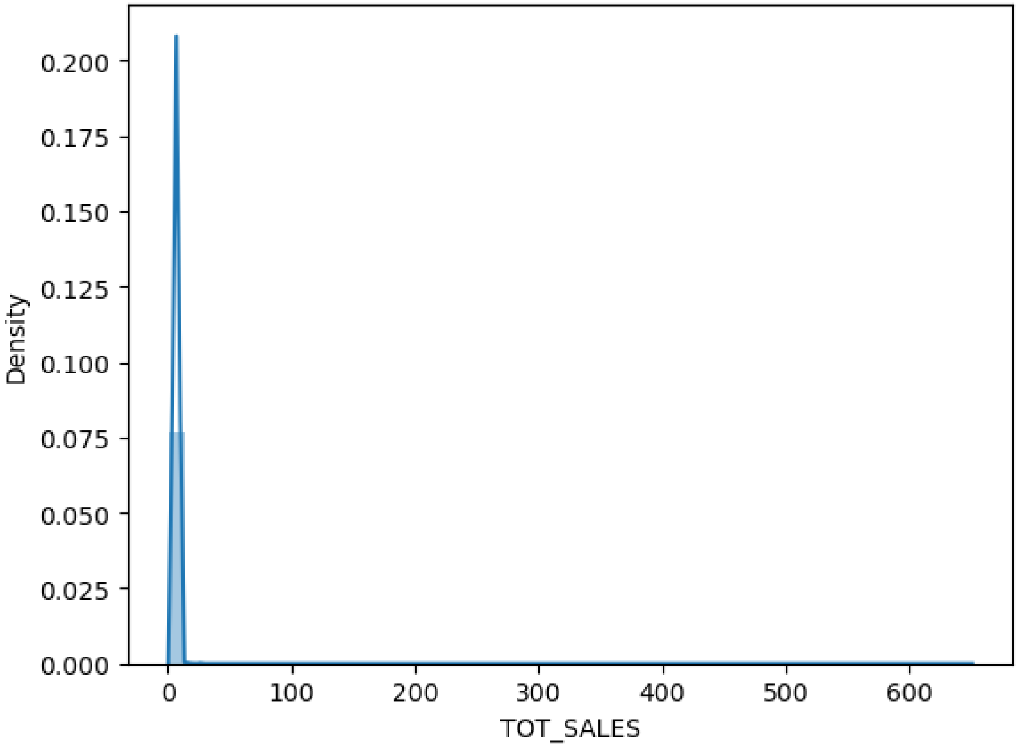
```
Out[35]: <Axes: xlabel='TOT_SALES'>
```



In [40]:
```
sns.distplot(transaction_data.TOT_SALES, kde = True)
```

```
C:\Users\Sreejith\AppData\Local\Temp\ipykernel_8988\3942976826.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751 (https://gist.github.co
m/mwaskom/de44147ed2974457ad6372750bbe5751)

  sns.distplot(transaction_data.TOT_SALES, kde = True)
```

Out[40]:
```
<Axes: xlabel='TOT_SALES', ylabel='Density'>
```



In [44]:
```
numeric_data = transaction_data.select_dtypes(['float','int'])
```

In [45]:
```
numeric_data
```

Out[45]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_SALES
0	43390	1	1000	1	5	2	6.0
1	43599	1	1307	348	66	3	6.3
2	43605	1	1343	383	61	2	2.9
3	43329	2	2373	974	69	5	15.0
4	43330	2	2426	1038	108	3	13.8
...	...	...	...	...	...	...	...
264831	43533	272	272319	270088	89	2	10.8
264832	43325	272	272358	270154	74	1	4.4
264833	43410	272	272379	270187	51	2	8.8
264834	43461	272	272379	270188	42	2	7.8
264835	43365	272	272380	270189	74	2	8.8

264836 rows × 7 columns

In [46]:
```
numeric_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264836 entries, 0 to 264835
Data columns (total 7 columns):
 #   Column          Non-Null Count   Dtype  
---  ------          --------------   -----  
 0   DATE            264836 non-null  int64  
 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_QTY        264836 non-null  int64  
 6   TOT_SALES       264836 non-null  float64
dtypes: float64(1), int64(6)
memory usage: 14.1 MB
```

In [47]:
```
numeric_data.describe()
```

Out[47]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_S
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	264836.000000	264836.000000	264836.0
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	56.583157	1.907309	7.3
std	105.389282	76.78418	8.057998e+04	7.813303e+04	32.826638	0.643654	3.0
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	1.000000	1.000000	1.5
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	28.000000	2.000000	5.4
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	56.000000	2.000000	7.4
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	85.000000	2.000000	9.2
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	114.000000	200.000000	650.0

In [52]:
```
x = numeric_data[numeric_data['TOT_SALES'] <8.0]
```

In [53]:
```
x
```

Out[53]:

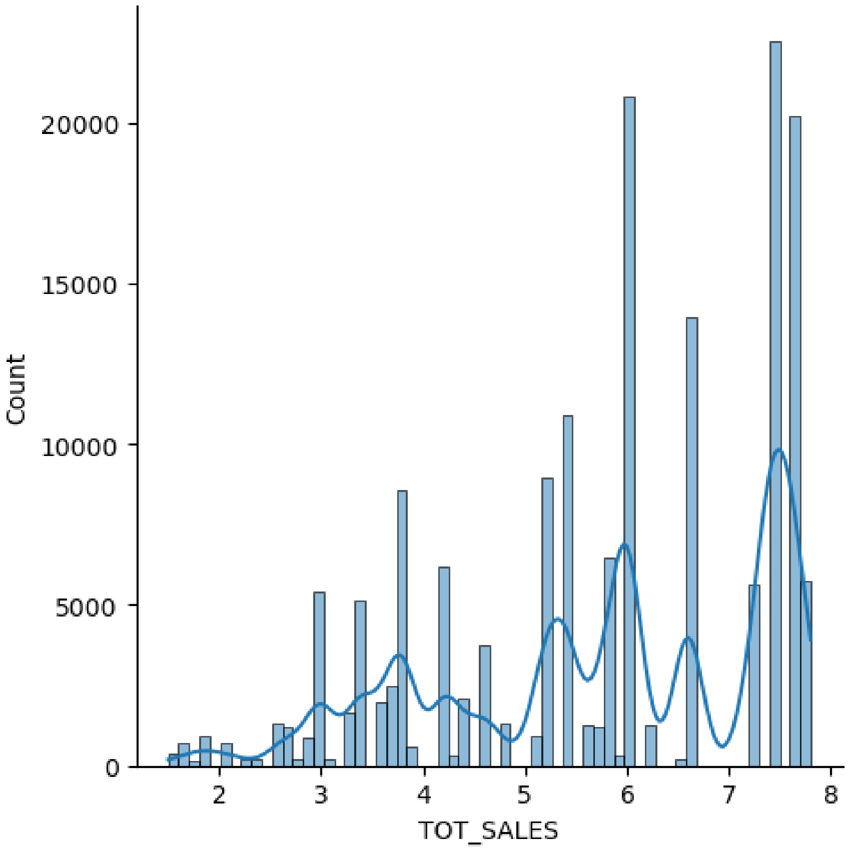
	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_SALES
0	43390	1	1000	1	5	2	6.0
1	43599	1	1307	348	66	3	6.3
2	43605	1	1343	383	61	2	2.9
5	43604	4	4074	2982	57	1	5.1
6	43601	4	4149	3333	16	1	5.7
...	...	...	...	...	...	...	...
264828	43308	272	272236	269974	68	2	7.4
264829	43540	272	272236	269976	49	2	7.6
264830	43416	272	272319	270087	44	2	6.6
264832	43325	272	272358	270154	74	1	4.4
264834	43461	272	272379	270188	42	2	7.8

166902 rows × 7 columns

In [55]:
```
sns.displot(x.TOT_SALES, kde = True)
```

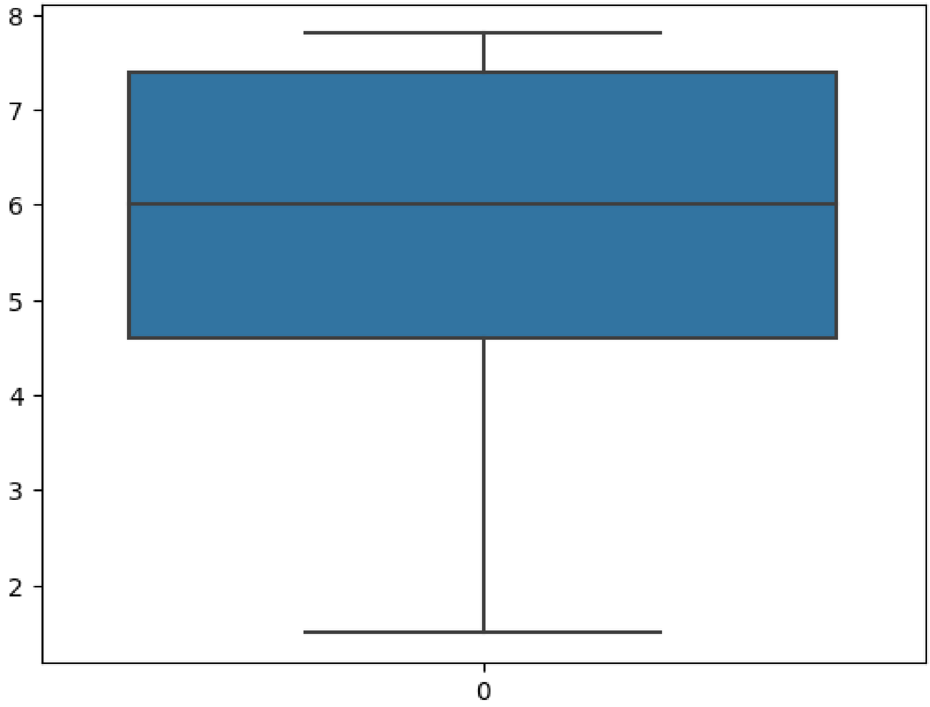
```
C:\Users\Sreejith\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The
figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[55]:
```
<seaborn.axisgrid.FacetGrid at 0x1bcb6b5f0d0>
```



In [56]:
```
sns.boxplot(x.TOT_SALES)
```

Out[56]:
```
<Axes: >
```



In [ ]:
```
```

Checking Data Formats

In [60]:
```
transaction_data
```

Out[60]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_NAME	PROD_QTY	TOT_SALE
0	43390	1	1000	1	5	Natural Chip Compny SeaSalt175g	2	6.
1	43599	1	1307	348	66	CCs Nacho Cheese 175g	3	6.
2	43605	1	1343	383	61	Smiths Crinkle Cut Chips Chicken 170g	2	2.
3	43329	2	2373	974	69	Smiths Chip Thinly S/Cream&Onion 175g	5	15.
4	43330	2	2426	1038	108	Kettle Tortilla ChpsHny&Jlpno Chili 150g	3	13.
...	...	...	...	...	...	...	...	.
264831	43533	272	272319	270088	89	Kettle Sweet Chilli And Sour Cream 175g	2	10.
264832	43325	272	272358	270154	74	Tostitos Splash Of Lime 175g	1	4.
264833	43410	272	272379	270187	51	Doritos Mexicana 170g	2	8.
264834	43461	272	272379	270188	42	Doritos Corn Chip Mexican Jalapeno 150g	2	7.
264835	43365	272	272380	270189	74	Tostitos Splash Of Lime 175g	2	8.

264836 rows × 8 columns

In [61]:
```
transaction_data.dtypes
```

Out[61]:
```
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
```

In [ ]:
```

```

In [ ]:
```

```