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

#import transaction dataset
transaction_data = pd.read_excel('QVI_transaction_data.xlsx')

transaction_data.head()
```

	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

[15]:

```
#import customer behaviour dataset
customer_data = pd.read_csv('QVI_purchase_behaviour.csv')
customer_data.head()
```

LYLTY_CARD_NBRLIFESTAGEPREMIUM_CUSTOMERO1000YOUNG

SINGLES/COUPLESPremium11002YOUNG SINGLES/COUPLESMainstream21003YOUNG FAMILIESBudget31004OLDER SINGLES/COUPLESMainstream41005MIDAGE SINGLES/COUPLESMainstream

SUMMARIZE THE DATASETS

transaction_data.describe()
DATESTORE_NBRLYLTY_CARD_NBRTXN_IDPROD_NBRPROD_QTYTOT_SALEScount264836.000000264836.000002.648360e+052.648360e+05264836.000000264836.000000264836.000000mean43464.036260135.080111.355495e+051.351583e+0556.5831571.9073097.304200std105.38928276.784188.057998e+047.813303e+0432.8266380.6436543.083226min43282.0000001.000001.000000e+031.000000e+001.000001.0000001.50000025%43373.00000070.000007.002100e+046.760150e+0428.0000002.0000005.40000050%43464.000000130.000001.303575e+051.351375e+0556.0000002.0000007.40000075%43555.000000203.000002.030942e+052.027012e+0585.0000002.0000009.200000max43646.000000272.000002.373711e+062.415841e+06114.000000200.000000650.000000

customer_data.describe()
LYLTY_CARD_NBRcount7.263700e+04mean1.361859e+05std8.989293e+04min1.000000e+0325%6.620200e+0450%1.340400e+0575%2.033750e+05max2.373711e+06

CHECK NULL

[19]:

```
transaction_data.isnull().sum()
```

```
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
```

```
transaction_data.dtypes
```

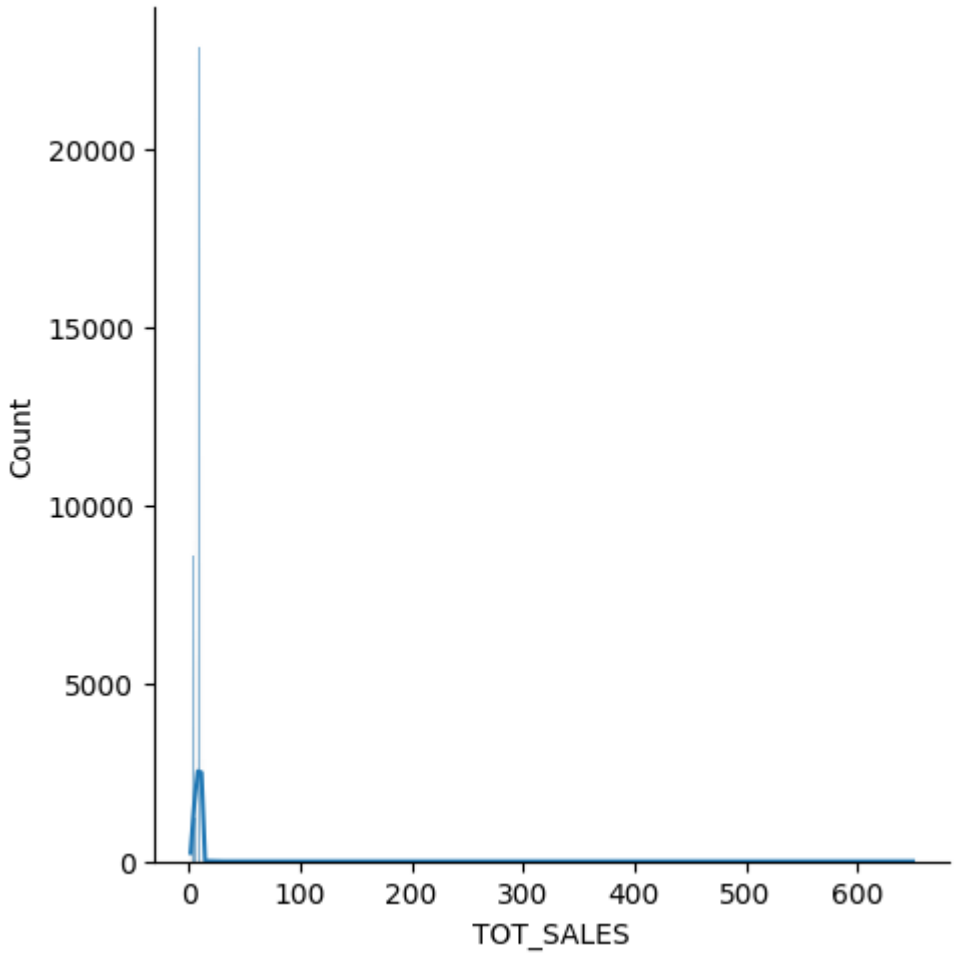
```
DATE              int64
STORE_NBR         int64
LYLTY_CARD_NBR    int64
TXN_ID            int64
PROD_NBR          int64
PROD_NAME        object
```

PROD_QTY int64
TOT_SALES float64
dtype: object

EXAMINE THE OUTLIERS

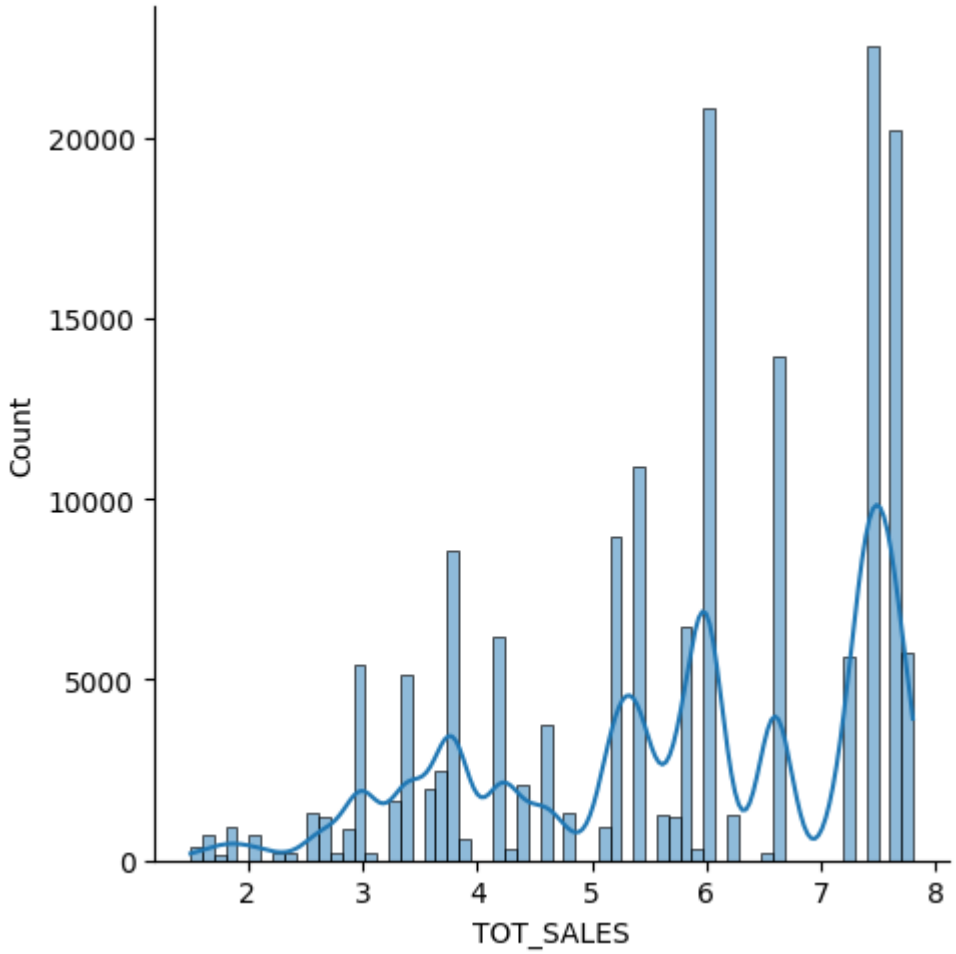
```
sns.displot(transaction_data.TOT_SALES, kde=True)
```

<seaborn.axisgrid.FacetGrid at 0x1ade4ad4a10>



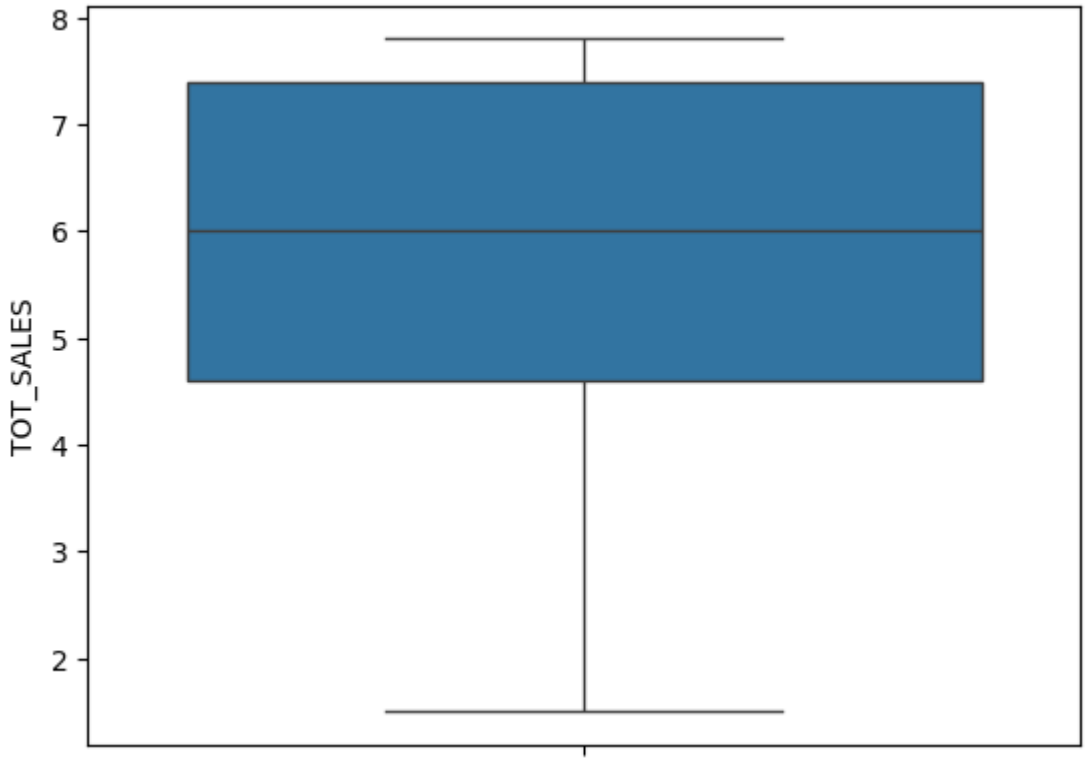

```
# checking in the float and int values for outliers
numeric_data = transaction_data.select_dtypes(['float', 'int'])
numeric_data.head()
```

<seaborn.axisgrid.FacetGrid at 0x1ade4d7b6e0>




```
# boxplot to show visually the outliers are present or not
sns.boxplot(x.TOT_SALES)
```

<Axes: ylabel='TOT_SALES'>



Therefore no Outlers a