In [6]:
```
#IMPORT REQUIRED LIBRARIES
```

In [1]:
```
import pandas as pd
import numpy as np
import seaborn as sns
```

In [5]:
```
pip install openpyxl
```

```
Defaulting to user installation because normal site-packages is not writeable
Collecting openpyxl
  Obtaining dependency information for openpyxl from https://files.pythonhosted.org/packages/58/d9/796181a30827b1210l786c21301f0f4536597a9249530916b1fdb5bbad91/openpyxl-3.1.3-py2.py3-none-any.whl.metadata
  Downloading openpyxl-3.1.3-py2.py3-none-any.whl.metadata (2.5 kB)
Collecting et-xmlfile (from openpyxl)
  Obtaining dependency information for et-xmlfile from https://files.pythonhosted.org/packages/96/c2/3dd434b0108730014f1b96fd286040dc3bcb70066346f7e01ec2ac95865f/et_xmlfile-1.1.0-py3-none-any.whl.metadata
  Downloading et_xmlfile-1.1.0-py3-none-any.whl.metadata (1.8 kB)
Downloading openpyxl-3.1.3-py2.py3-none-any.whl (251 kB)
   ---------------------------------------- 0.0/251.3 kB ? eta -:--:--
   ---- ----------------------------------- 30.7/251.3 kB 660.6 kB/s eta 0:00:01
   ---------- ----------------------------- 245.8/251.3 kB 3.0 MB/s eta 0:00:01
   ---------------------------------------- 251.3/251.3 kB 3.1 MB/s eta 0:00:00
Downloading et_xmlfile-1.1.0-py3-none-any.whl (4.7 kB)
Installing collected packages: et-xmlfile, openpyxl
Successfully installed et-xmlfile-1.1.0 openpyxl-3.1.3
Note: you may need to restart the kernel to use updated packages.
[notice] A new release of pip is available: 23.2.1 -> 24.0
[notice] To update, run: python.exe -m pip install --upgrade pip
```

In [6]:
```
dataset = pd.read_excel('QVI_transaction_data.xlsx')
```

In [8]:
```
dataset.head()
```

Out[8]:

	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 [1]:
```
#SUMMARIZE DATASET
```

In [9]:
```
dataset.describe()
```

Out[9]:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	PROD_NBR	PROD_QTY	TOT_SALES
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	264836.000000	264836.000000	264836.000000
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	56.583157	1.907309	7.304200
std	105.389282	76.78418	8.057998e+04	7.813303e+04	32.826638	0.643654	3.083226
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	1.000000	1.000000	1.500000
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	28.000000	2.000000	5.400000
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	56.000000	2.000000	7.400000
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	85.000000	2.000000	9.200000
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	114.000000	200.000000	650.000000

In [2]:
```
#CHECKING FOR NULL VALUES
```

In [10]:
```
dataset.isnull().sum()
```

Out[10]:
```
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]:
```
dataset2 = pd.read_csv('QVI_purchase_behaviour.csv')
```

In [18]:
```
dataset2.head()
```

Out[18]:

	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 [3]:
```
#SUMMARIZE DATASET2
```

In [19]:
```
dataset2.describe()
```

Out[19]:

	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 [4]:
```
#CHECKING FOR NULL VALUES
```

In [20]:
```
dataset2.isnull().sum()
```

Out[20]:
```
LYLTY_CARD_NBR      0
LIFESTAGE           0
PREMIUM_CUSTOMER    0
dtype: int64
```

In [21]:
```
data_types = dataset.dtypes
print(data_types)
```

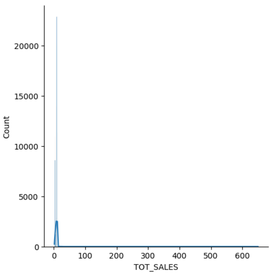
```
DATE              int64
STORE_NBR         int64
LYLTY_CARD_NBR    int64
TXN_ID            int64
PROD_NBR          int64
PROD_NAME        object
PROD_QTY          int64
TOT_SALES       float64
dtype: object
```

In [5]:
```
#EXAMINE THE OUTLIERS
```

In [22]:
```
import matplotlib.pyplot as plt
```

In [23]:
```
sns.displot(dataset.TOT_SALES, kde = True)
```

Out[23]:
```
<seaborn.axisgrid.FacetGrid at 0x2af995e5610>
```



In [24]:
```
numericdata = dataset.select_dtypes(['float', 'int'])
numericdata.head()
```

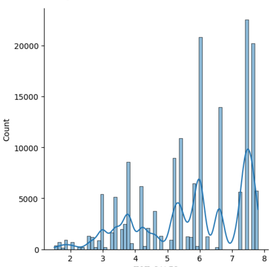
Out[24]:

	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

In [25]:
```
x = numericdata[numericdata['TOT_SALES']<8.000]
```

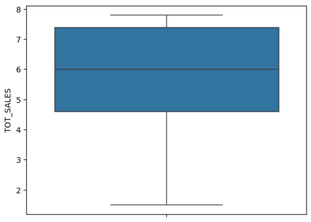
In [26]:
```
sns.displot(x.TOT_SALES, kde = True)
```

Out[26]:
```
<seaborn.axisgrid.FacetGrid at 0x2af995e5550>
```



In [27]:
```
sns.boxplot(x.TOT_SALES)
```

Out[27]:
```
<Axes: ylabel='TOT_SALES'>
```


TOT_SALES
8
7
6
5
4
3
2