Code

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

# Load the data

data = pd.read_csv('QVI_transaction_data.csv')

# Step 1: Data Overview

print("Data Overview:")

print(data.info())  # Basic information about the dataset

# Step 2: Check for Missing Values

print("\nMissing Values:")

print(data.isnull().sum())  # Count the missing values in each column

# Step 3: Data Summary Statistics

print("\nSummary Statistics:")

print(data.describe())  # Summary statistics for numeric columns

# Step 4: Data Visualization (using Matplotlib)

# Box Plot for 'TXN_ID', 'PROD_QTY', and 'TOT_SALES'

plt.figure(figsize=(4, 4))

plt.boxplot([data['TXN_ID'], data['PROD_QTY'], data['TOT_SALES']], labels=['TXN_ID', 'PROD_QTY',
'TOT_SALES'])

plt.xlabel('Columns')

plt.ylabel('Values')

plt.title('Box Plot')

plt.grid(True)

plt.show()

# Histogram for 'TOT_SALES'

plt.figure(figsize=(8, 5))

plt.hist(data['TOT_SALES'], bins=20, color='skyblue', edgecolor='black', density=True)
```

```
plt.xlabel('TOT_SALES')

plt.ylabel('Density')

plt.title('Histogram')

plt.grid(True)

plt.show()

# Scatter Plot for 'PROD_QTY' vs. 'TOT_SALES' with the specified range

plt.figure(figsize=(8, 5))

plt.scatter(data['PROD_QTY'], data['TOT_SALES'], color='orange', alpha=0.7)

plt.xlabel('PROD_QTY')

plt.ylabel('TOT_SALES')

plt.title('Scatter Plot')

plt.grid(True)

# Set the range for the x-axis and y-axis

plt.xlim(0, 8)

plt.ylim(0, 50)

plt.show()
```

Output

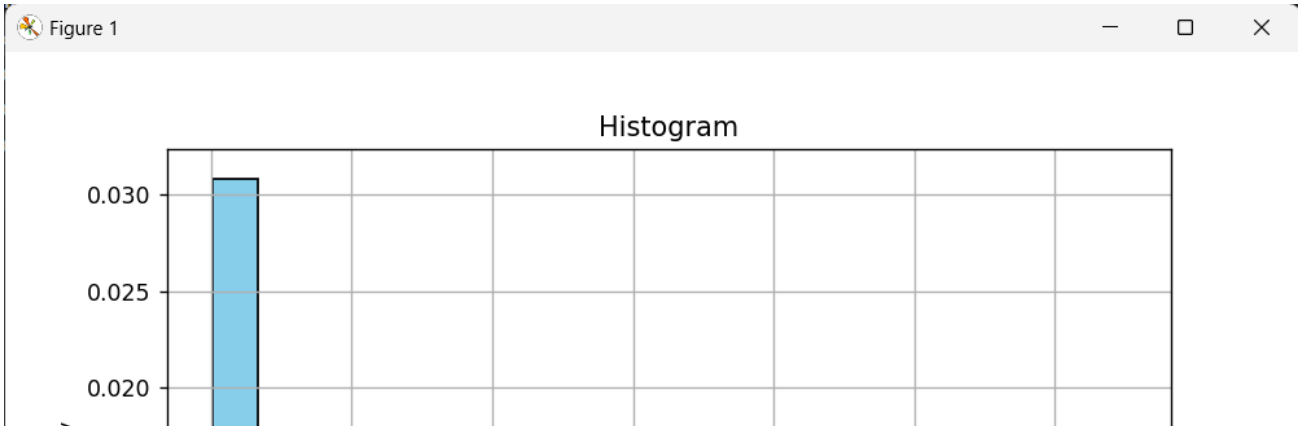
```
Data Overview:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 264836 entries, 0 to 264835
Data columns (total 8 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   DATE            264836 non-null  int64
 1   STORE_NBR       264836 non-null  int64
 2   LYLTY_CARD_NBR  264836 non-null  int64
 3   TXN_ID          264836 non-null  int64
 4   PROD_NBR        264836 non-null  int64
 5   PROD_NAME       264836 non-null  object
 6   PROD_QTY        264836 non-null  int64
 7   TOT_SALES       264836 non-null  float64
dtypes: float64(1), int64(6), object(1)
memory usage: 16.2+ MB
None
```

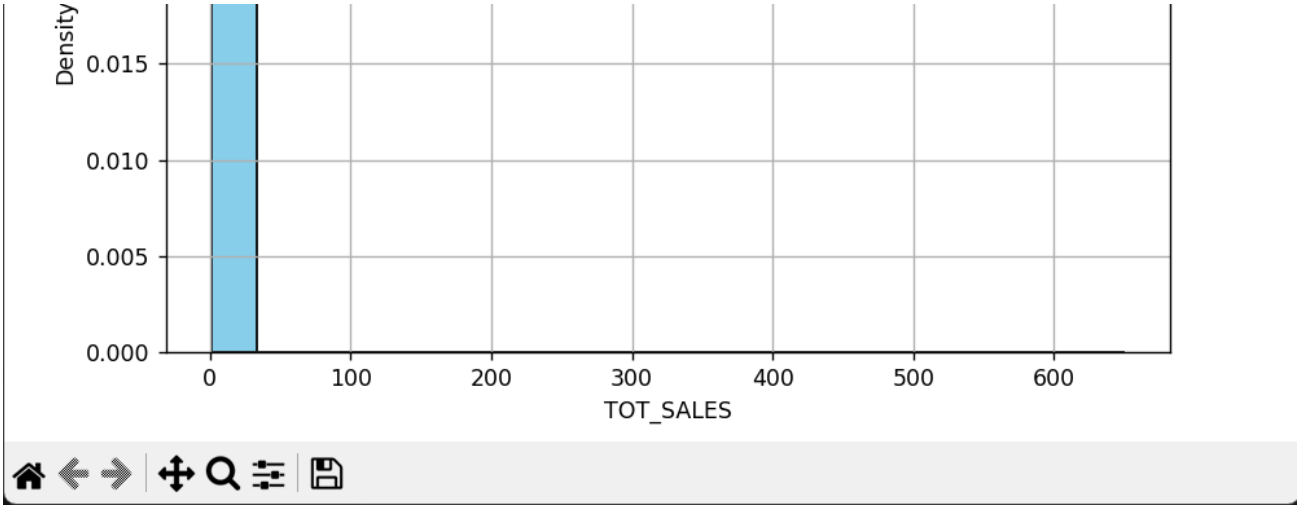
```
Missing Values:
DATE              0
STORE_NBR         0
LYLTY_CARD_NBR    0
TXN_ID            0
PROD_NBR          0
PROD_NAME         0
PROD_QTY          0
TOT_SALES         0
dtype: int64
```

Summary Statistics:

	DATE	STORE_NBR	LYLTY_CARD_NBR	TXN_ID	\
count	264836.000000	264836.00000	2.648360e+05	2.648360e+05	
mean	43464.036260	135.08011	1.355495e+05	1.351583e+05	
std	105.389282	76.78418	8.057998e+04	7.813303e+04	
min	43282.000000	1.00000	1.000000e+03	1.000000e+00	
25%	43373.000000	70.00000	7.002100e+04	6.760150e+04	
50%	43464.000000	130.00000	1.303575e+05	1.351375e+05	
75%	43555.000000	203.00000	2.030942e+05	2.027012e+05	
max	43646.000000	272.00000	2.373711e+06	2.415841e+06	

	PROD_NBR	PROD_QTY	TOT_SALES
count	264836.000000	264836.000000	264836.000000
mean	56.583157	1.907309	7.304200
std	32.826638	0.643654	3.083226
min	1.000000	1.000000	1.500000
25%	28.000000	2.000000	5.400000
50%	56.000000	2.000000	7.400000
75%	85.000000	2.000000	9.200000
max	114.000000	200.000000	650.000000




Density
0.015
0.010
0.005
0.000
0
100
200
300
400
500
600
TOT_SALES

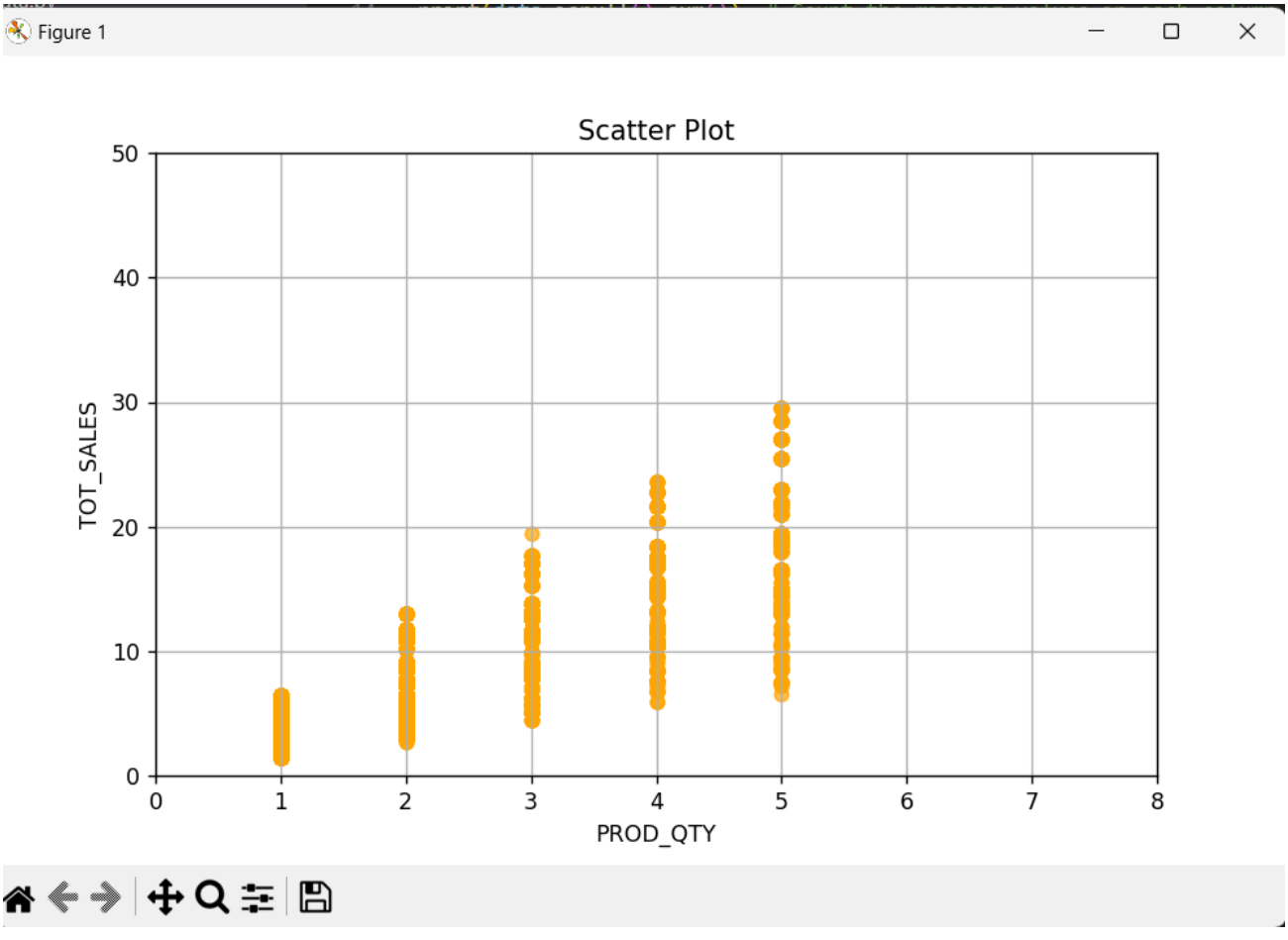


Figure 1
Scatter Plot
TOT_SALES
PROD_QTY
0
10
20
30
40
50
0
1
2
3
4
5
6
7
8

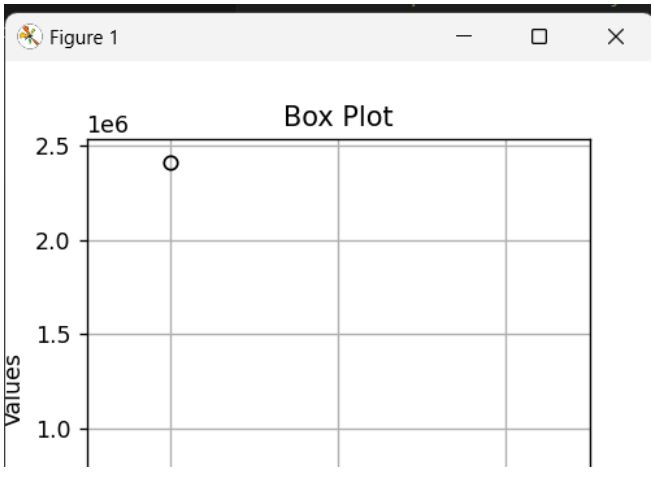
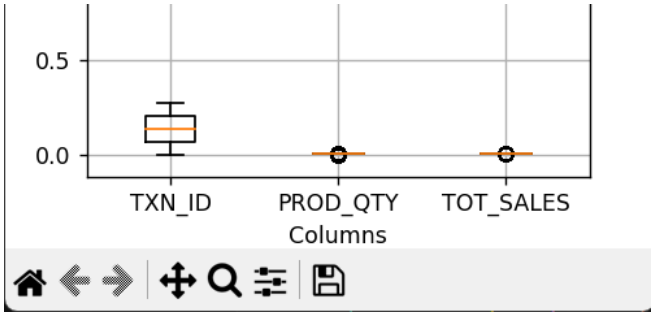


Figure 1
Box Plot
1e6
Values
2.5
2.0
1.5
1.0


0.5
0.0
TXN_ID
PROD_QTY
TOT_SALES
Columns