In [1]:
```
# import the libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
# load the dataset
df=pd.read_csv('D:/Downloads/UM/Financial Analytics data.csv')
df
```

Out[2]:

	S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore	Unnamed: 4
0	1	Reliance Inds.	583436.72	99810.00	NaN
1	2	TCS	563709.84	30904.00	NaN
2	3	HDFC Bank	482953.59	20581.27	NaN
3	4	ITC	320985.27	9772.02	NaN
4	5	H D F C	289497.37	16840.51	NaN
...	...	...	...	...	...
483	496	Lak. Vilas Bank	3029.57	790.17	NaN
484	497	NOCIL	3026.26	249.27	NaN
485	498	Orient Cement	3024.32	511.53	NaN
486	499	Natl.Fertilizer	3017.07	2840.75	NaN
487	500	L T Foods	NaN	NaN	NaN

488 rows × 5 columns

In [3]:
```
# check the brief info of the data
df.head()
```

Out[3]:

	S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore	Unnamed: 4
0	1	Reliance Inds.	583436.72	99810.00	NaN
1	2	TCS	563709.84	30904.00	NaN
2	3	HDFC Bank	482953.59	20581.27	NaN
3	4	ITC	320985.27	9772.02	NaN
4	5	H D F C	289497.37	16840.51	NaN

In [4]:
```
# shape of the data
df.shape
```

Out[4]:
```
(488, 5)
```

In [5]:
```
# index of the dataset
df.index
```

Out[5]:
```
RangeIndex(start=0, stop=488, step=1)
```

In [7]:
```
# columns
df.columns
```

Out[7]:
```
Index(['S.No.', 'Name', 'Mar Cap - Crore', 'Sales Qtr - Crore', 'Unnamed: 4'], dtype='object')
```

In [8]:
```
# value counts of name field
df['Name'].value_counts()
```

Out[8]:
```
Reliance Inds.      1
Dishman Carbogen    1
Timken India        1
GE Power            1
Guj Alkalies        1
                   ..
Tata Global         1
Reliance Nip.Lif    1
Apollo Hospitals    1
MphasiS             1
L T Foods           1
Name: Name, Length: 488, dtype: int64
```

In [9]:
```
# index of the dataset
df.index
```

Out[9]:
```
RangeIndex(start=0, stop=488, step=1)
```

In [10]:
```
# columns of the dataset
df.columns
```

Out[10]:
```
Index(['S.No.', 'Name', 'Mar Cap - Crore', 'Sales Qtr - Crore', 'Unnamed: 4'], dtype='object')
```

In [11]:
```
# add the missing values from unnamed column to sales qtr-crore
df['Sales Qtr - Crore'].fillna(df['Unnamed: 4'],inplace=True)
```

In [12]:
```
df
```

Out[12]:

	S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore	Unnamed: 4
0	1	Reliance Inds.	583436.72	99810.00	NaN
1	2	TCS	563709.84	30904.00	NaN
2	3	HDFC Bank	482953.59	20581.27	NaN
3	4	ITC	320985.27	9772.02	NaN
4	5	H D F C	289497.37	16840.51	NaN
...	...	...	...	...	...
483	496	Lak. Vilas Bank	3029.57	790.17	NaN
484	497	NOCIL	3026.26	249.27	NaN
485	498	Orient Cement	3024.32	511.53	NaN
486	499	Natl.Fertilizer	3017.07	2840.75	NaN
487	500	L T Foods	NaN	NaN	NaN

488 rows × 5 columns

In [13]:
```
# remove the unnamed column
df.drop(columns='Unnamed: 4',axis=1,inplace=True)
```

In [14]:
```
df
```

Out[14]:

	S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore
0	1	Reliance Inds.	583436.72	99810.00
1	2	TCS	563709.84	30904.00
2	3	HDFC Bank	482953.59	20581.27
3	4	ITC	320985.27	9772.02
4	5	H D F C	289497.37	16840.51
...	...	...	...	...
483	496	Lak. Vilas Bank	3029.57	790.17
484	497	NOCIL	3026.26	249.27
485	498	Orient Cement	3024.32	511.53
486	499	Natl.Fertilizer	3017.07	2840.75
487	500	L T Foods	NaN	NaN

488 rows × 4 columns

In [15]:
```
# check the null values in a dataset
df.isnull().sum()
```

Out[15]:
```
S.No.                0
Name                 0
Mar Cap - Crore      9
Sales Qtr - Crore    29
dtype: int64
```

In [16]:
```
#  check if the duplicated values present in the dataset
df.duplicated().sum()
```

Out[16]:
```
0
```

In [17]:
```
# check the skewness of the mar cap-crore column
df['Mar Cap - Crore'].skew()
```

Out[17]:
```
5.560197674089212
```

In [18]:
```
# # fill the null values of Mar cap feature with median value there is an outliers so we have to fill with medi
df['Mar Cap - Crore'].fillna(df['Mar Cap - Crore'].median(),inplace=True)
```

In [19]:
```
# check the skewness of the sales qtr-crore column
df['Sales Qtr - Crore'].skew()
```

Out[19]:
```
6.833822665838169
```

In [20]:
```
# fill the null values of sales qtr feature with median value there is an outliers so we have to fill with medi
df['Sales Qtr - Crore'].fillna(df['Sales Qtr - Crore'].median(),inplace=True)
```

In [21]:
```
# after filling the values with median there is no null values
df.isnull().sum()
```

Out[21]:
```
S.No.                0
Name                 0
Mar Cap - Crore      0
Sales Qtr - Crore    0
dtype: int64
```

In [22]:
```
# copy the cleaned dataset into new dataset
df1=df.copy()
```

In [23]:
```
df1
```

Out[23]:

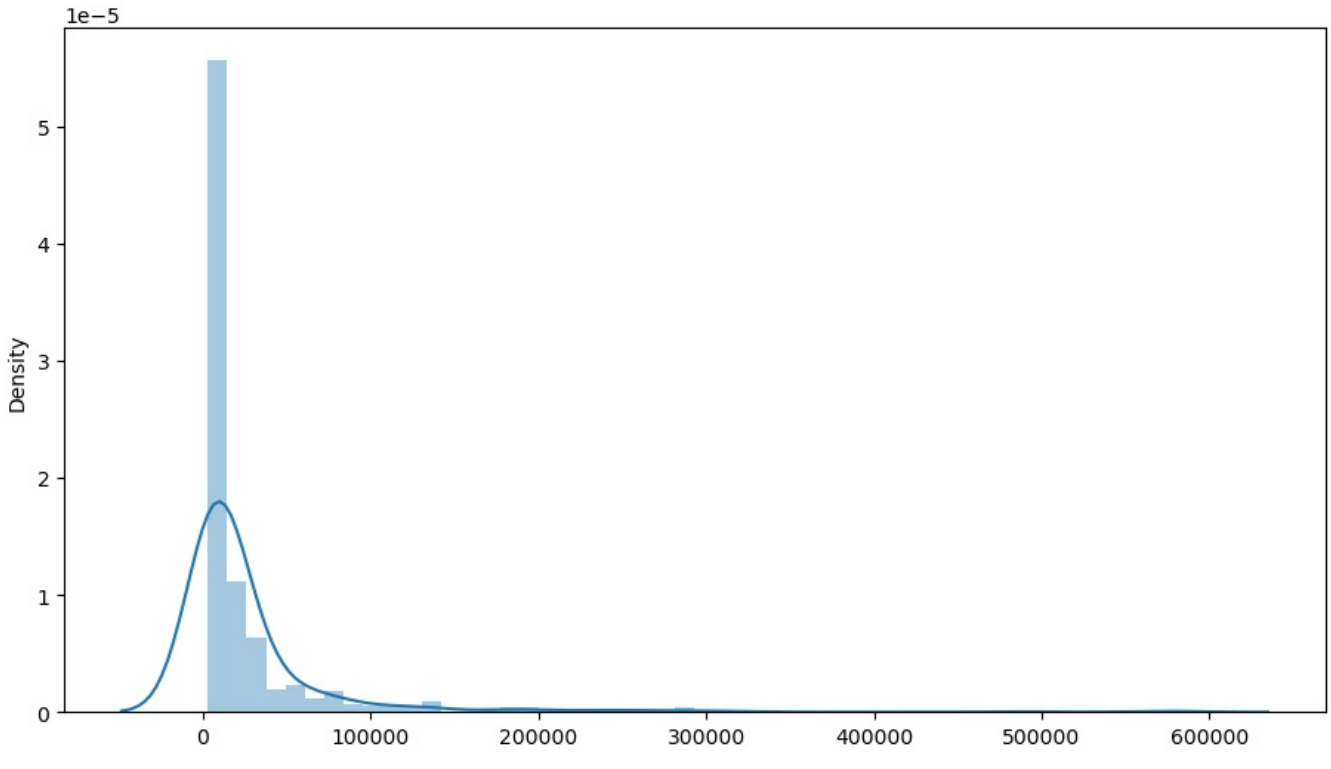
	S.No.	Name	Mar Cap - Crore	Sales Qtr - Crore
0	1	Reliance Inds.	583436.72	99810.00
1	2	TCS	563709.84	30904.00
2	3	HDFC Bank	482953.59	20581.27
3	4	ITC	320985.27	9772.02
4	5	H D F C	289497.37	16840.51
...	...	...	...	...
483	496	Lak. Vilas Bank	3029.57	790.17
484	497	NOCIL	3026.26	249.27
485	498	Orient Cement	3024.32	511.53
486	499	Natl.Fertilizer	3017.07	2840.75
487	500	L T Foods	9885.05	1137.17

488 rows × 4 columns

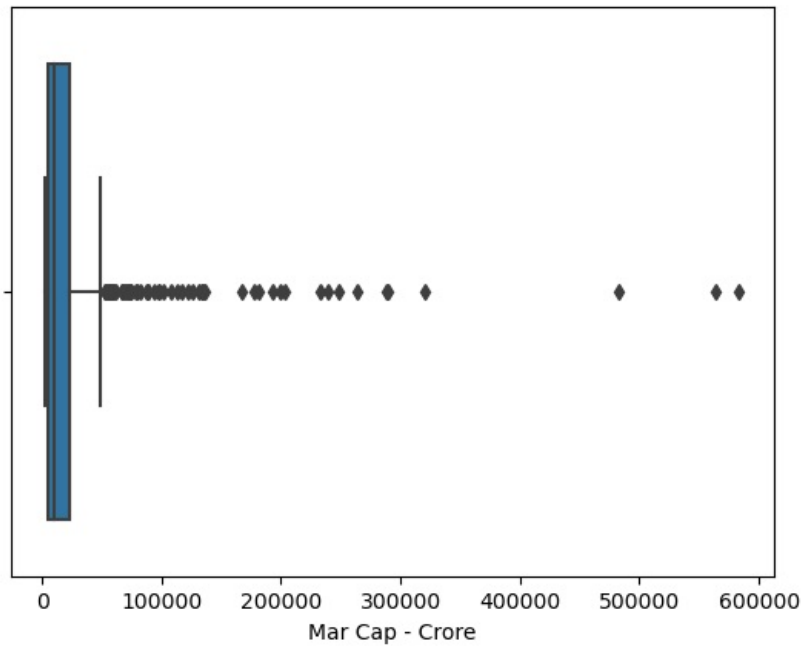
In [24]:
```
# check the null values of the new dataset
df1.isnull().sum()
```

Out[24]:
```
S.No.                0
Name                 0
Mar Cap - Crore      0
Sales Qtr - Crore    0
dtype: int64
```

In [25]:
```
# check the distribution of Mar cap column
plt.figure(figsize=(11,6))
sns.distplot(x=df['Mar Cap - Crore'])
plt.show()
```



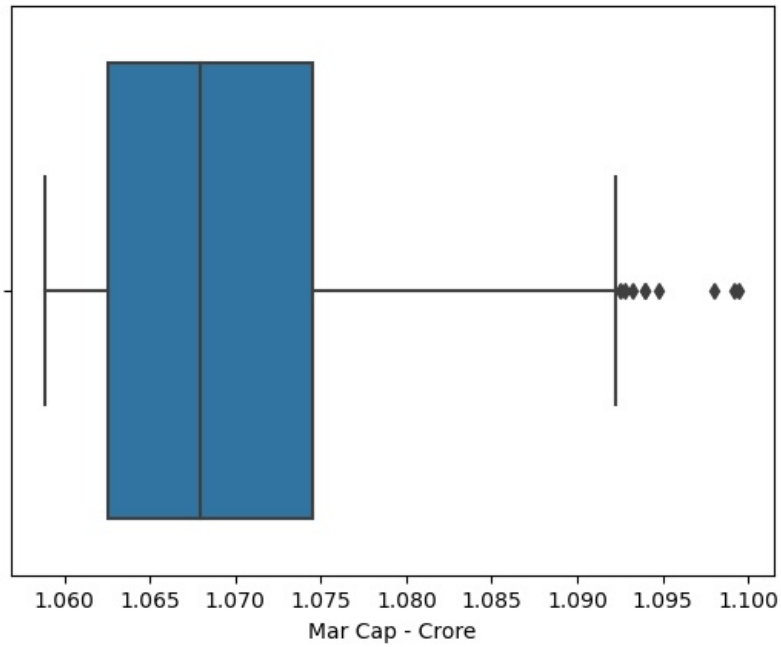
In [26]:
```
# check the outliers of Mar cap column
sns.boxplot(x=df['Mar Cap - Crore'])
plt.show()
```



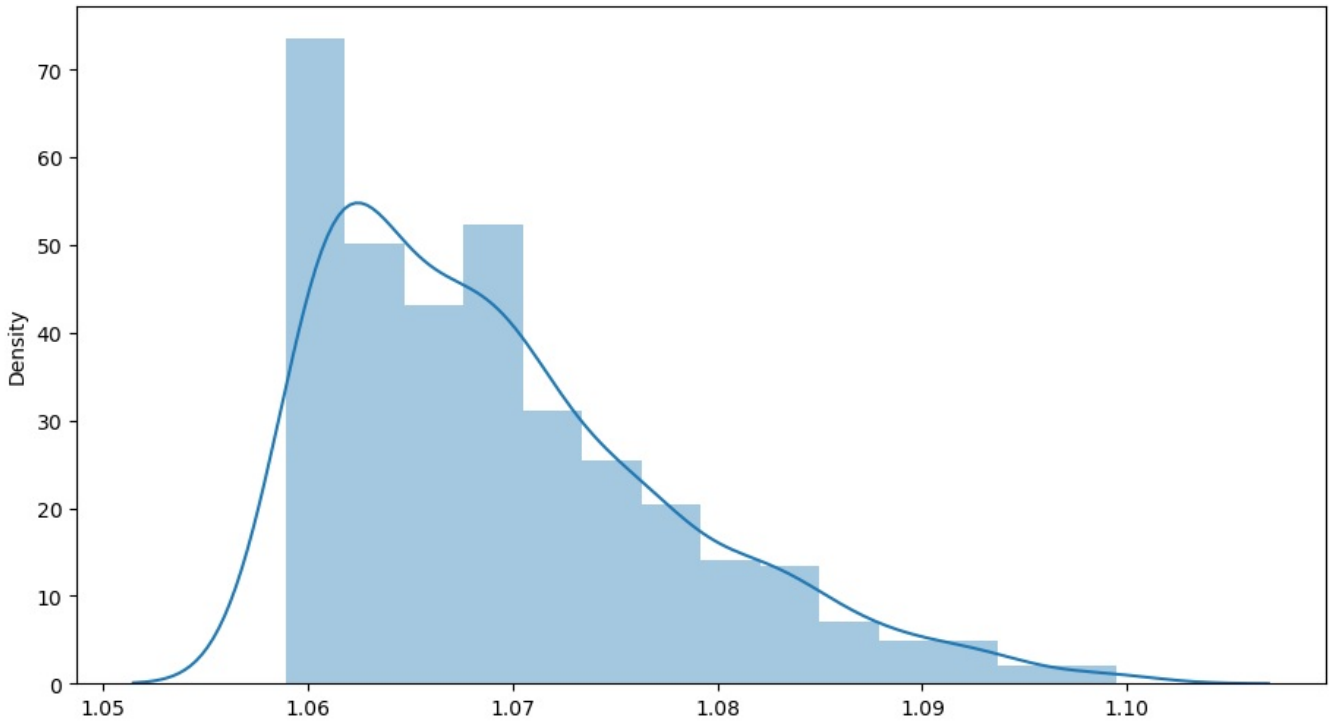
In [28]:
```
# remove the outliers by using root transformation
df['Mar Cap - Crore']=df['Mar Cap - Crore']**(1/115)
df['Mar Cap - Crore'].skew()
```

Out[28]:
```
0.9937104105455516
```

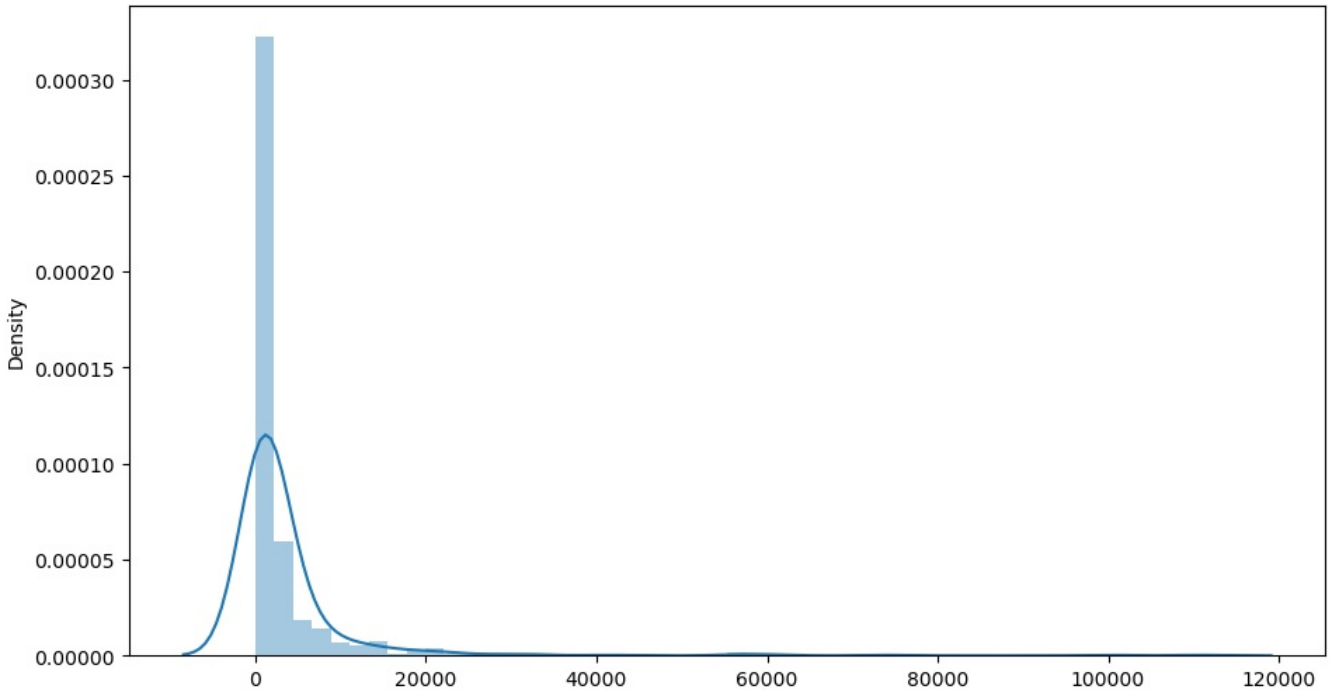
In [21]:
```
# after removing the outliers sill there is an outliers
sns.boxplot(x=df['Mar Cap - Crore'])
plt.show()
```



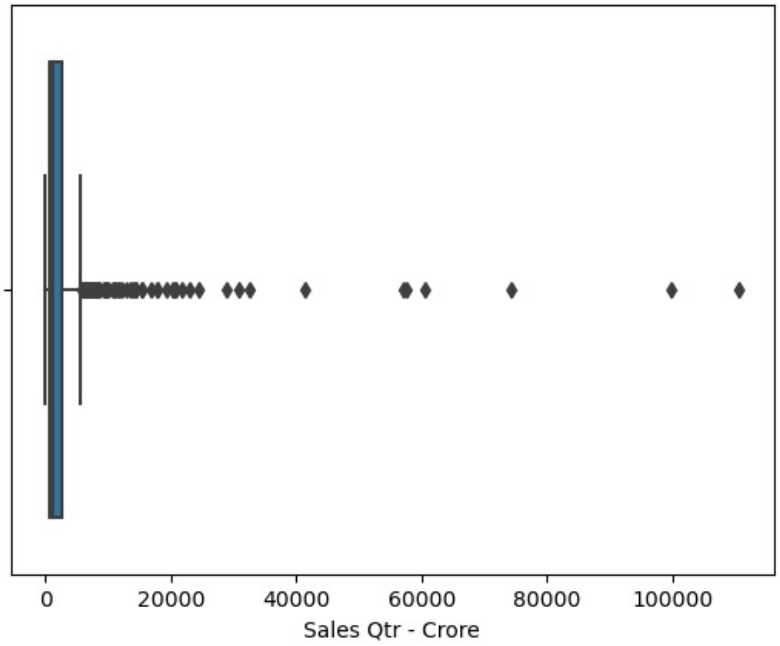
In [22]:
```
# distribution after removel of outliers
plt.figure(figsize=(11,6))
sns.distplot(x=df['Mar Cap - Crore'])
plt.show()
```



In [23]:
```
# ccheck the distribution of sales qtr feature
plt.figure(figsize=(11,6))
sns.distplot(x=df['Sales Qtr - Crore'])
plt.show()
```



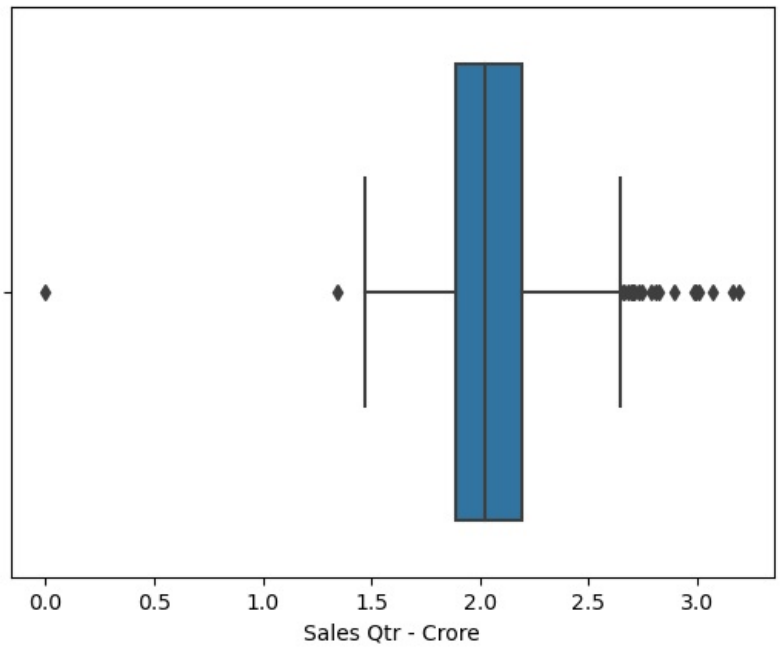
In [24]:
```
# check the outliers of sales qtr column
sns.boxplot(x=df['Sales Qtr - Crore'])
plt.show()
```



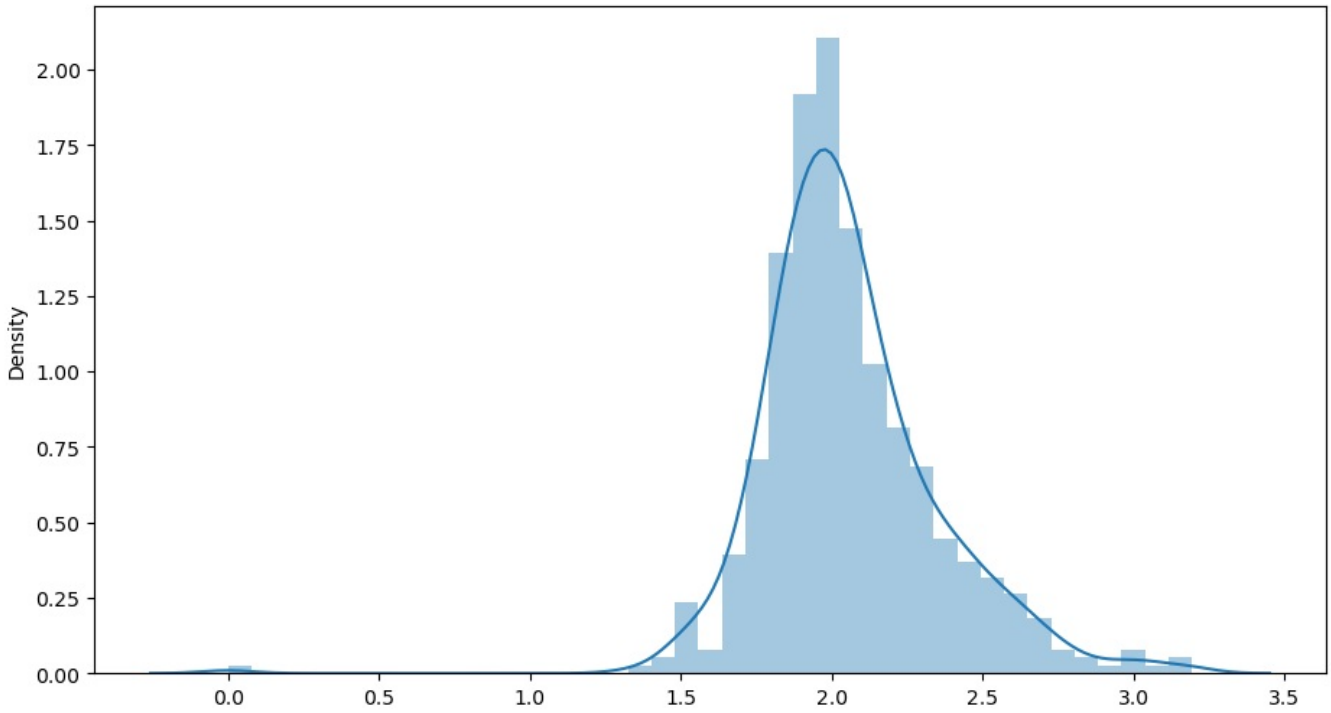
In [25]:
```
# remove the outliers by using root transformation
df['Sales Qtr - Crore']=df['Sales Qtr - Crore']**(1/10)
df['Sales Qtr - Crore'].skew()
```

Out[25]:
```
0.13832236638944503
```

In [26]:
```
# after removing the outliers sill there is an outliers
sns.boxplot(x=df['Sales Qtr - Crore'])
plt.show()
```



In [27]:
```
# distribution after removel of outliers
plt.figure(figsize=(11,6))
sns.distplot(x=df['Sales Qtr - Crore'])
plt.show()
```



In [29]:
```
# convert into new data file
df1.to_csv('Finamne_Data.csv',index=False)
```

Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js