In [13]:
```
import pandas as pd
```

In [14]:
```
df = pd.read_csv('xAPI-Edu-Data.csv')
```

In [15]:
```
df.head(5)
```

Out[15]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relatio
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe

In [5]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64 
 10  VisITedResources          480 non-null    int64 
 11  AnnouncementsView         480 non-null    int64 
 12  Discussion                480 non-null    int64 
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

In [6]:
```
df.isnull().sum()
```

Out[6]:
```
gender                      0
NationalITy                 0
PlaceofBirth                0
StageID                     0
GradeID                     0
SectionID                   0
Topic                       0
Semester                    0
Relation                    0
raisedhands                 0
VisITedResources            0
AnnouncementsView           0
Discussion                  0
ParentAnsweringSurvey       0
ParentschoolSatisfaction    0
StudentAbsenceDays          0
Class                       0
dtype: int64
```

In [7]:
```
import numpy as np
new_df = df['raisedhands'].replace(np.nan,0)
```

In [8]:
```
new_df.isnull().sum()
```

Out[8]:
```
0
```

In [9]:
```
df.dropna(axis=0,inplace=True)
```

In [10]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 480 entries, 0 to 479
Data columns (total 17 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   gender                    480 non-null    object
 1   NationalITy               480 non-null    object
 2   PlaceofBirth              480 non-null    object
 3   StageID                   480 non-null    object
 4   GradeID                   480 non-null    object
 5   SectionID                 480 non-null    object
 6   Topic                     480 non-null    object
 7   Semester                  480 non-null    object
 8   Relation                  480 non-null    object
 9   raisedhands               480 non-null    int64 
 10  VisITedResources          480 non-null    int64 
 11  AnnouncementsView         480 non-null    int64 
 12  Discussion                480 non-null    int64 
 13  ParentAnsweringSurvey     480 non-null    object
 14  ParentschoolSatisfaction  480 non-null    object
 15  StudentAbsenceDays        480 non-null    object
 16  Class                     480 non-null    object
dtypes: int64(4), object(13)
memory usage: 63.9+ KB
```

In [11]:

```
import seaborn as sb
import warnings
import matplotlib.pyplot as plt
```

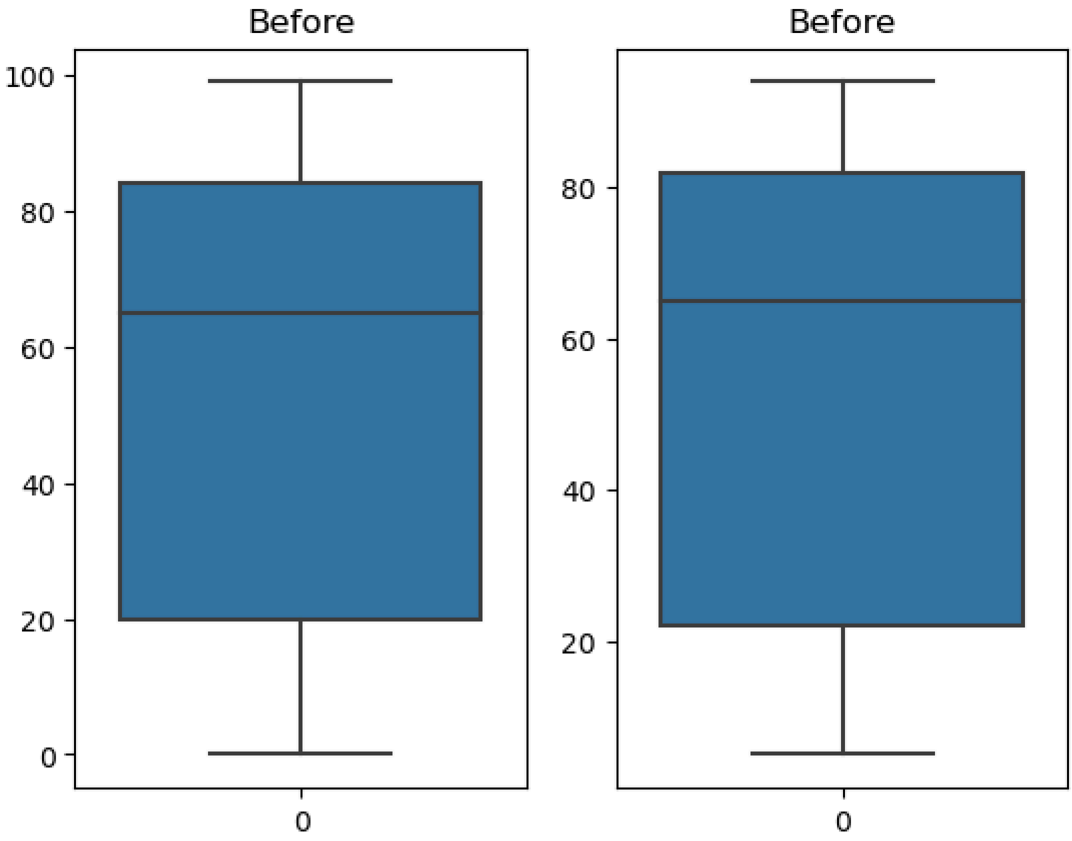
In [23]:

```
warnings.filterwarnings("ignore")
fig,axis = plt.subplots(1,2)
max_val = df.VisITedResources.quantile(0.95)
min_val = df.VisITedResources.quantile(0.05)
print("Before Shape",df.shape)
df2 = df[(df['VisITedResources']>min_val) & (df['VisITedResources']<max_val
print("After Shape",df2.shape)
sb.boxplot(df['VisITedResources'],orient='v',ax=axis[0])
axis[0].title.set_text("Before")
sb.boxplot(df2['VisITedResources'],orient='v',ax=axis[1])
axis[1].title.set_text("Before")
plt.show
```

```
Before Shape (480, 17)
After Shape (427, 17)
```

Out[23]:

```
<function matplotlib.pyplot.show(close=None, block=None)>
```



In [24]:

```
df.head()
```

Out[24]:

	gender	NationalITy	PlaceofBirth	StageID	GradeID	SectionID	Topic	Semester	Relatio
0	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
1	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
2	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
3	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe
4	M	KW	KuwaIT	lowerlevel	G-04	A	IT	F	Fathe

In [29]:

```
from sklearn.preprocessing import StandardScaler
```

In [31]:

```
scaler = StandardScaler()
x = df[['raisedhands','VisITedResources','AnnouncementsView','Discussion']]
scaledf = scaler.fit_transform(x)
print(scaledf)
```

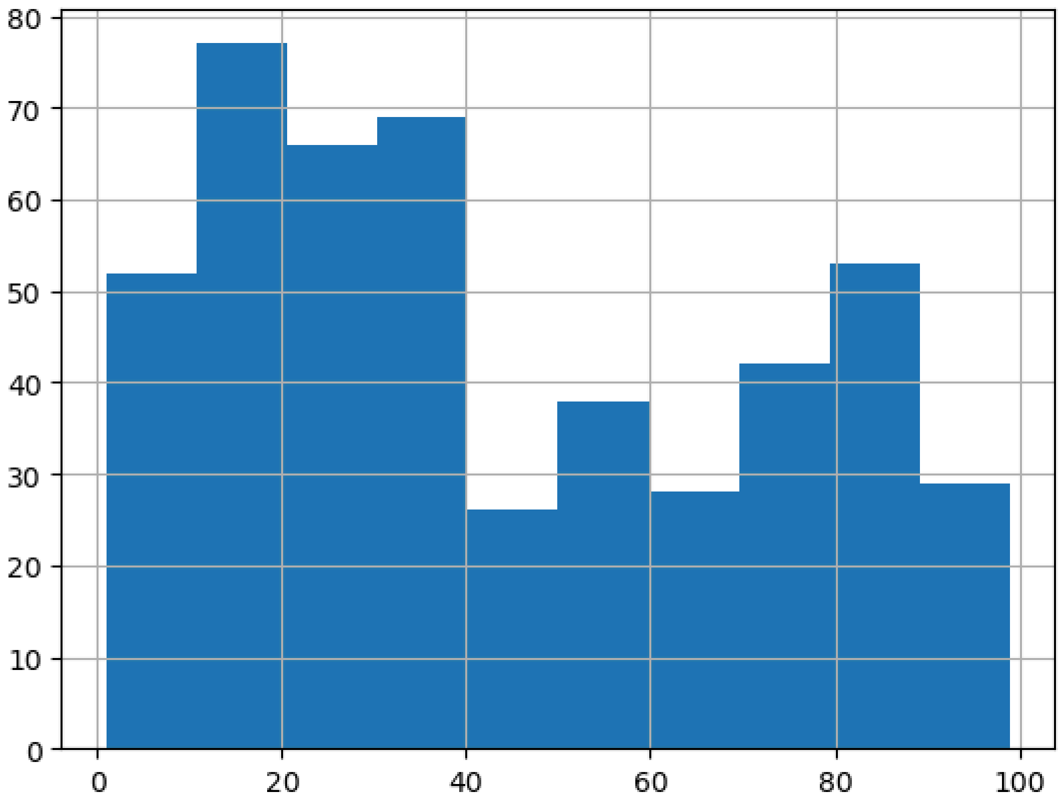
```
[[-1.03342931 -1.17407456 -1.35116659 -0.84332615]
 [-0.87081258 -1.05302945 -1.31354928 -0.66222533]
 [-1.19604604 -1.44642607 -1.4264012  -0.48112451]
 ...
 [ 0.26750452  0.58107959 -0.48596856 -0.51734468]
 [-0.54557912 -1.14381328 -0.89975892  0.49681992]
 [-0.3829624  -1.23459712 -0.56120318  0.67792074]]
```

In [32]:

```
df.Discussion.hist()
```

Out[32]:

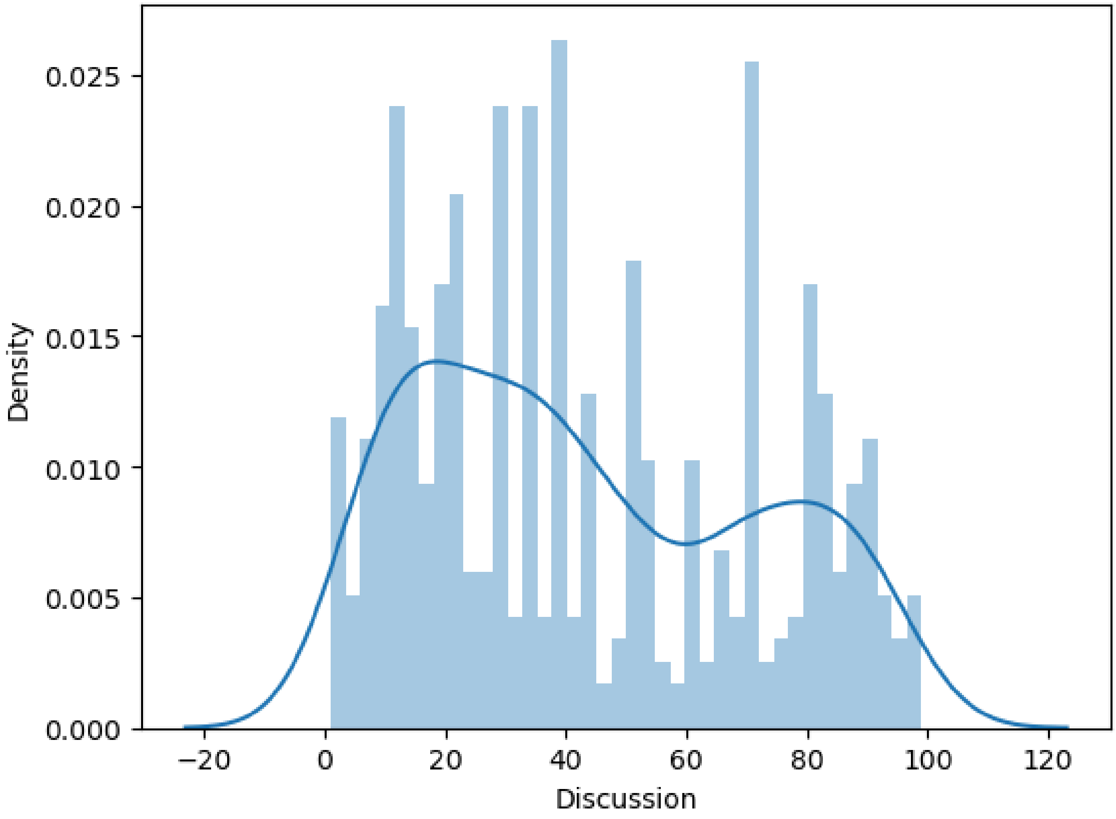
```
<Axes: >
```



```
In [33]: import scipy.stats as stats
```

```
In [35]: sb.distplot(df['Discussion'],bins=40)
```

```
Out[35]: <Axes: xlabel='Discussion', ylabel='Density'>
```



```
In [36]: df['Discussion'].skew()
```

```
Out[36]: 0.3625939845015566
```

```
In [ ]:
```