17. Function (Transformer)

- to conver the non-normal distribution data into normal distribution data

In [18]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pylab as plt
import numpy as np
```

In [3]:
```
dataset = pd.read_csv('loan.csv')
```

In [4]:
```
dataset.head(3)
```

Out[4]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	C
0	LP001002	Male	No	0	Graduate	No	5849	
1	LP001003	Male	Yes	1	Graduate	No	4583	
2	LP001005	Male	Yes	0	Graduate	Yes	3000	

In [7]:
```
dataset.isnull().sum()
```

Out[7]:
```
Loan_ID               0
Gender               13
Married               3
Dependents           15
Education             0
Self_Employed        32
ApplicantIncome       0
CoapplicantIncome     0
LoanAmount           22
Loan_Amount_Term     14
Credit_History       50
Property_Area         0
Loan_Status           0
dtype: int64
```

In [8]:
```
sns.distplot(dataset['CoapplicantIncome'])
plt.show()
```

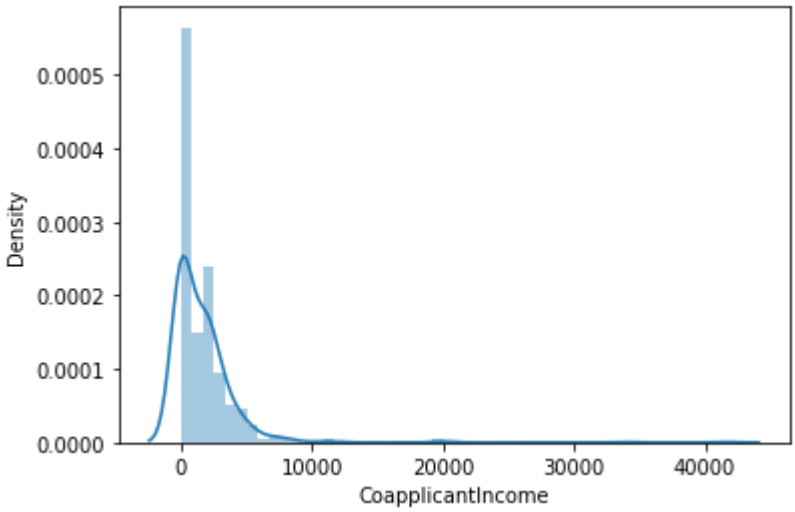
```
C:\Users\rashi\AppData\Local\Temp/ipykernel_4868/3783729653.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dataset['CoapplicantIncome'])
```



- You can see the data is not normally distributed as it has long tail on right side

17.1 Remove Outlier

We will remove the outlier by IQR method

In [9]:
```
q1 = dataset['CoapplicantIncome'].quantile(0.25)
q3 = dataset['CoapplicantIncome'].quantile(0.75)
iqr = q3 - q1
```

In [12]:
```
min_r = q1-(1.5*iqr)
max_r = q3 + (1.5*iqr)
min_r, max_r
```

Out[12]:
```
(-3445.875, 5743.125)
```

In [15]:
```
dataset = dataset[dataset['CoapplicantIncome'] <= max_r]
dataset
```

Out[15]:

	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome
0	LP001002	Male	No	0	Graduate	No	5849
1	LP001003	Male	Yes	1	Graduate	No	4583
2	LP001005	Male	Yes	0	Graduate	Yes	3000
3	LP001006	Male	Yes	0	Not Graduate	No	2583
4	LP001008	Male	No	0	Graduate	No	6000
...	...	...	...	...	...	...	...
609	LP002978	Female	No	0	Graduate	No	2900
610	LP002979	Male	Yes	3+	Graduate	No	4106
611	LP002983	Male	Yes	1	Graduate	No	8072
612	LP002984	Male	Yes	2	Graduate	No	7583
613	LP002990	Female	No	0	Graduate	Yes	4583

596 rows × 13 columns

In [16]:

```
sns.distplot(dataset['CoapplicantIncome'])
plt.show()
```

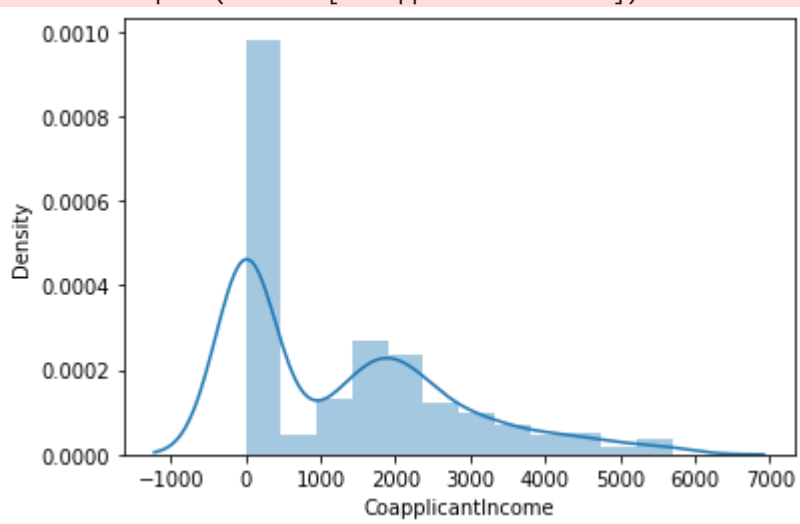
```
C:\Users\rashi\AppData\Local\Temp/ipykernel_4868/3783729653.py:1: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dataset['CoapplicantIncome'])
```



17.2 Function Transformation

- To make the data normally distributed, we will use function transformation

In [17]:
```
from sklearn.preprocessing import FunctionTransformer
```

In [21]:
```
ft = FunctionTransformer(func=np.log1p)
```

In [22]:
```
ft.fit(dataset[['CoapplicantIncome']])
```

Out[22]:
```
▾ FunctionTransformer
FunctionTransformer(func=<ufunc 'log1p'>)
```

In [25]:
```
dataset['CoapplicantIncome_tf'] = ft.transform(dataset[['CoapplicantIncome']])
dataset['CoapplicantIncome_tf']
```

Out[25]:
```
0      0.000000
1      7.319202
2      0.000000
3      7.765993
4      0.000000
         ...
609    0.000000
610    0.000000
611    5.484797
612    0.000000
613    0.000000
Name: CoapplicantIncome_tf, Length: 596, dtype: float64
```

In [28]:
```
plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
sns.distplot(dataset['CoapplicantIncome'])
plt.title("Before")
plt.subplot(1,2,2)
sns.distplot(dataset['CoapplicantIncome_tf'])
plt.title("After")
plt.show()
```

```
C:\Users\rashi\AppData\Local\Temp/ipykernel_4868/3310440801.py:3: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dataset['CoapplicantIncome'])
C:\Users\rashi\AppData\Local\Temp/ipykernel_4868/3310440801.py:6: UserWarning:

`distplot` is a deprecated function and will be removed in seaborn v0.14.0.

Please adapt your code to use either `displot` (a figure-level function with
similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see
https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751

  sns.distplot(dataset['CoapplicantIncome_tf'])
```