Chapter 5 - Ex1: Mammals

Cho dữ liệu mammals.csv chứa thông tin về mammals.

- Phân tích thông tin sơ bộ về dữ liệu trên hai thuộc tính BrainWt, BodyWt, xem xét mối quan hệ của 2 thuộc tính này. Trực quan hóa dữ liệu.
- Để dự đoán BrainWt dựa trên BodyWt cần phải kiểm tra và chuẩn hóa dữ liệu. Hãy chọn một phương pháp để chuẩn hóa dữ liệu dựa trên thông tin nêu trên. Trực quan hóa dữ liệu sau khi chuẩn hóa.

In [1]:

```
import pandas as pd
import numpy as np
```

In [2]:

```
data = pd.read_csv("mammals.csv", index_col=0)
data.head()
```

Out[2]:

	Species	BodyWt	BrainWt	NonDreaming	Dreaming	TotalSleep	LifeSpan	G
1	Africanelephant	6654.000	5712.0	NaN	NaN	3.3	38.6	
2	Africangiantpouchedrat	1.000	6.6	6.3	2.0	8.3	4.5	
3	ArcticFox	3.385	44.5	NaN	NaN	12.5	14.0	
4	Arcticgroundsquirrel	0.920	5.7	NaN	NaN	16.5	NaN	
5	Asianelephant	2547.000	4603.0	2.1	1.8	3.9	69.0	

In [3]:

```
data.describe()
```

Out[3]:

	BodyWt	BrainWt	NonDreaming	Dreaming	TotalSleep	LifeSpan	Gestatio
count	62.000000	62.000000	48.000000	50.000000	58.000000	58.000000	58.00000
mean	198.789984	283.134194	8.672917	1.972000	10.532759	19.877586	142.35344
std	899.158011	930.278942	3.666452	1.442651	4.606760	18.206255	146.80503
min	0.005000	0.140000	2.100000	0.000000	2.600000	2.000000	12.00000
25%	0.600000	4.250000	6.250000	0.900000	8.050000	6.625000	35.75000
50%	3.342500	17.250000	8.350000	1.800000	10.450000	15.100000	79.00000
75%	48.202500	166.000000	11.000000	2.550000	13.200000	27.750000	207.50000
max	6654.000000	5712.000000	17.900000	6.600000	19.900000	100.000000	645.00000

In [4]:

```
body_range = data.BodyWt.ptp()
body_range
```

Out[4]:

```
6653.995
```

In [5]:

```
brain_range = data.BrainWt.ptp()
brain_range
```

Out[5]:

```
5711.86
```

In [6]:

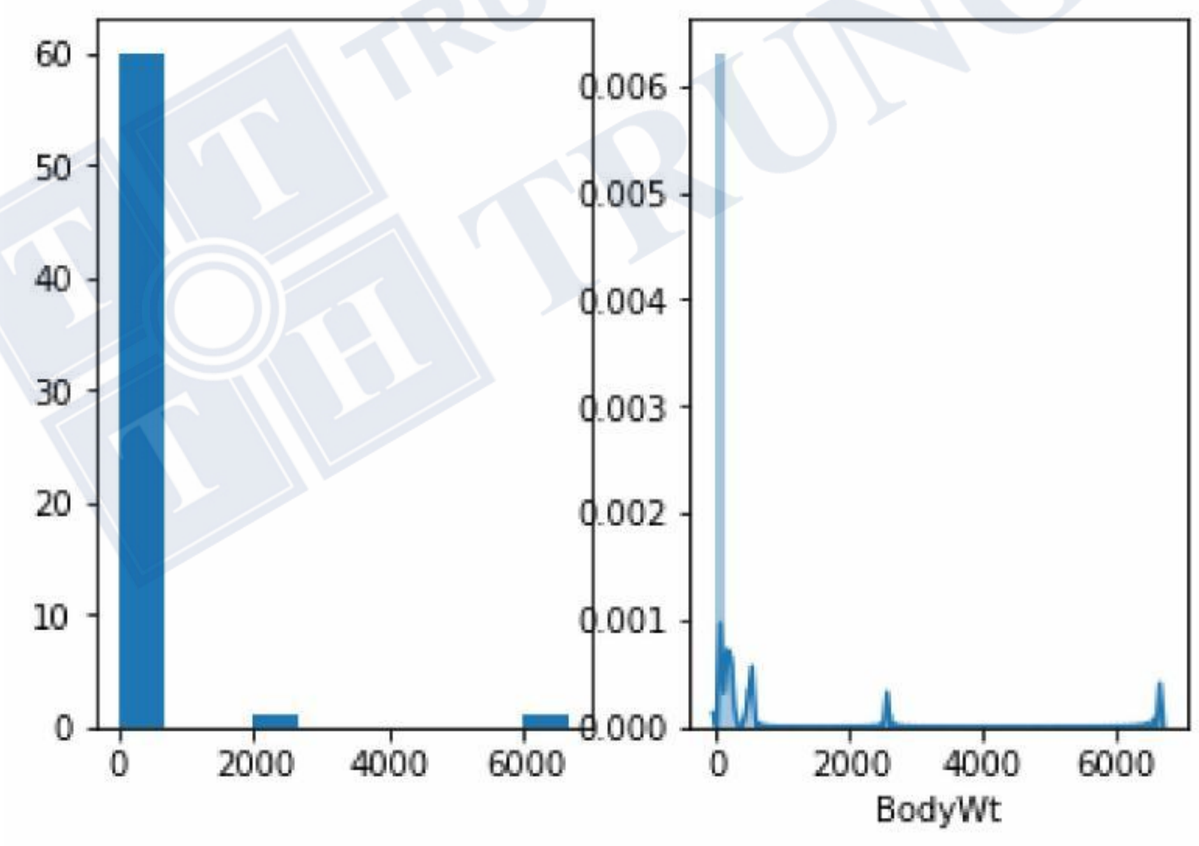
```
# Có một khoảng lớn giữa min và max
```

In [7]:

```
import matplotlib.pyplot as plt
import seaborn as sns
```

In [8]:

```
plt.subplot(1,2,1)
plt.hist(data.BodyWt)
plt.subplot(1,2,2)
sns.distplot(data.BodyWt)
plt.show()
```



In [9]:

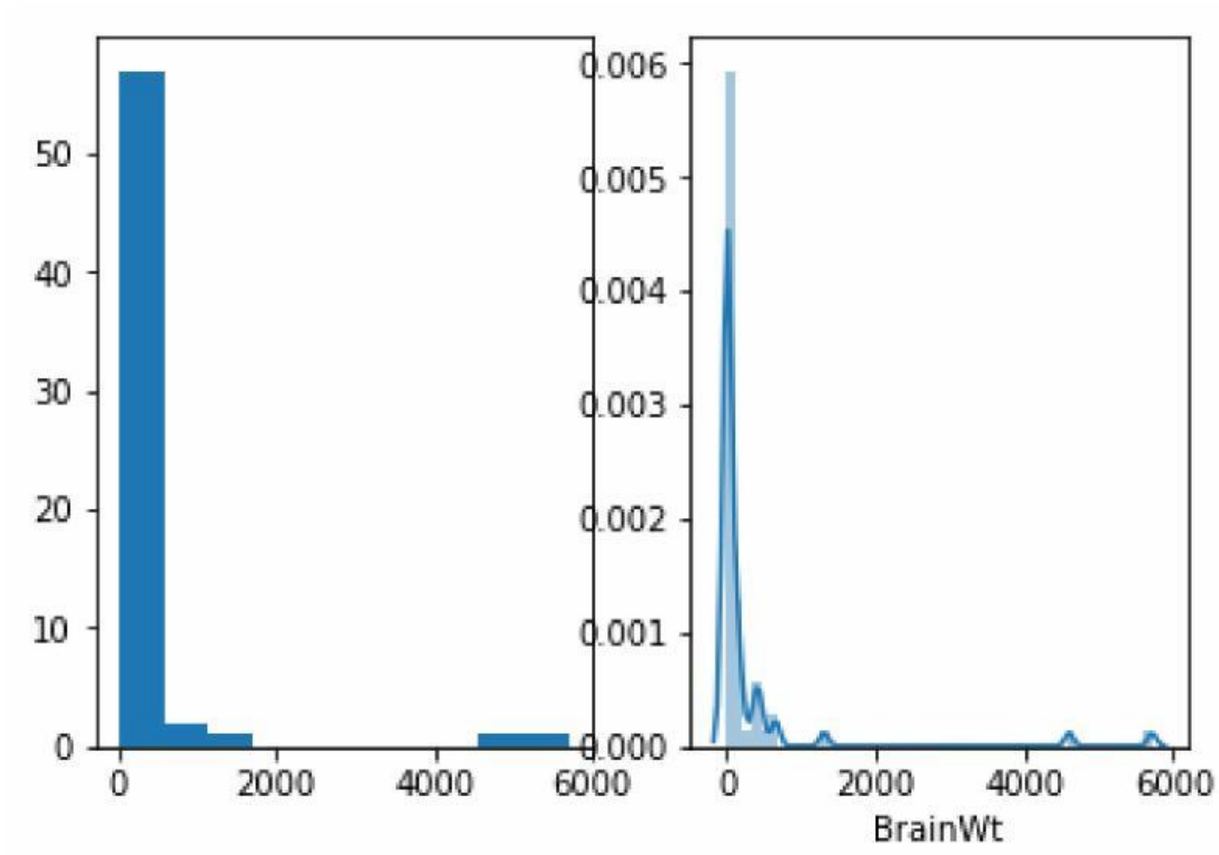
```
data.BodyWt.skew() # phân phối lệch phải
```

Out[9]:

```
6.563608062833757
```

In [10]:

```
plt.subplot(1,2,1)
plt.hist(data.BrainWt)
plt.subplot(1,2,2)
sns.distplot(data.BrainWt)
plt.show()
```



In [11]:

```
data.BrainWt.skew() # phân phối lệch phải
```

Out[11]:

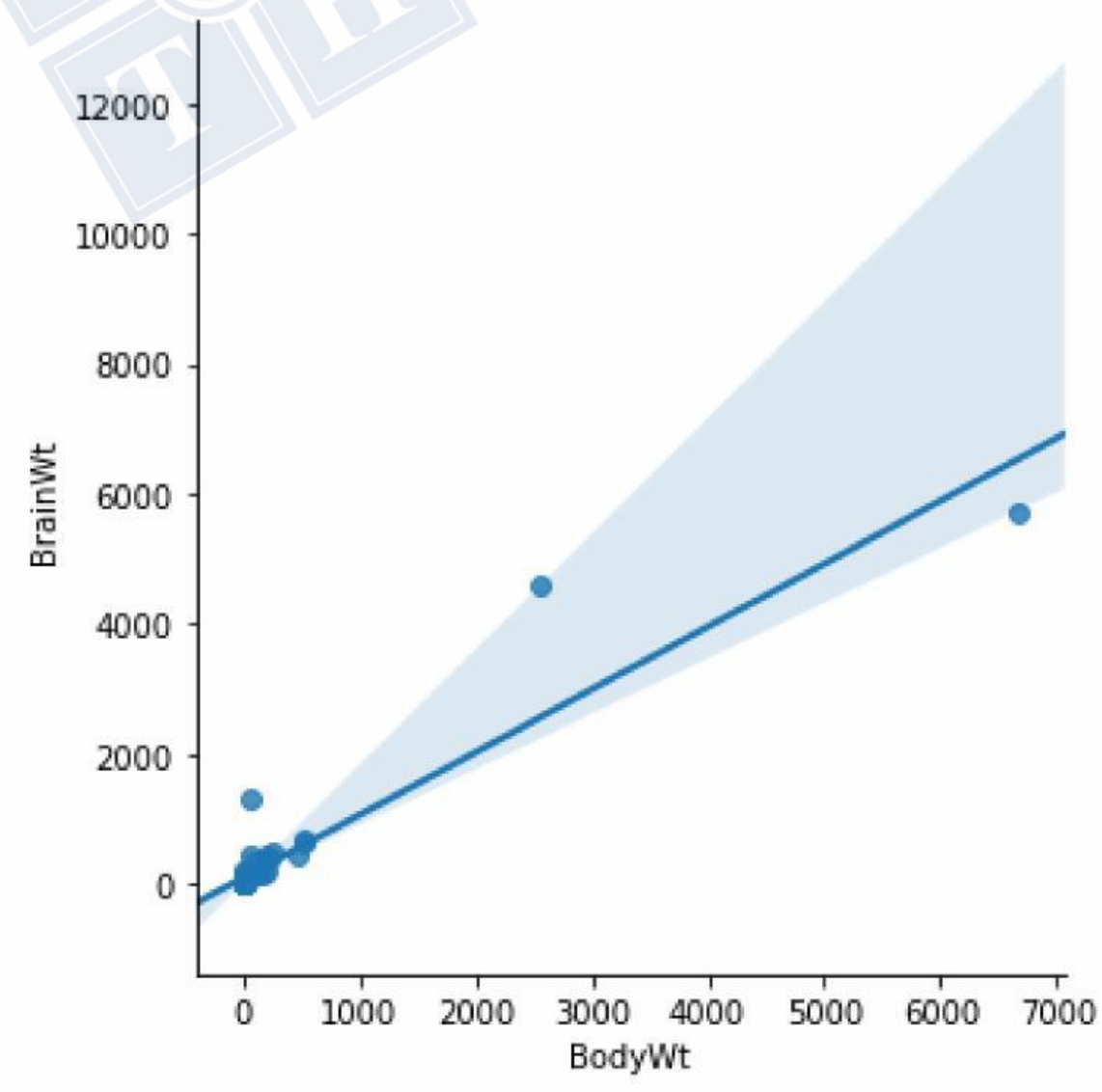
```
5.071589456939673
```

In [12]:

```
# Xem xét mối quan hệ
sns.lmplot(data=data, x='BodyWt', y='BrainWt')
```

Out[12]:

```
<seaborn.axisgrid.FacetGrid at 0x21520687e48>
```



In [13]:

```
# Quan hệ tuyến tính
# Chọn phương pháp chuẩn hóa là log normalization
```

In [14]:

```
data['BodyWt_log'] = np.log(data.BodyWt)
data['BrainWt_log'] = np.log(data.BrainWt)
```

In [15]:

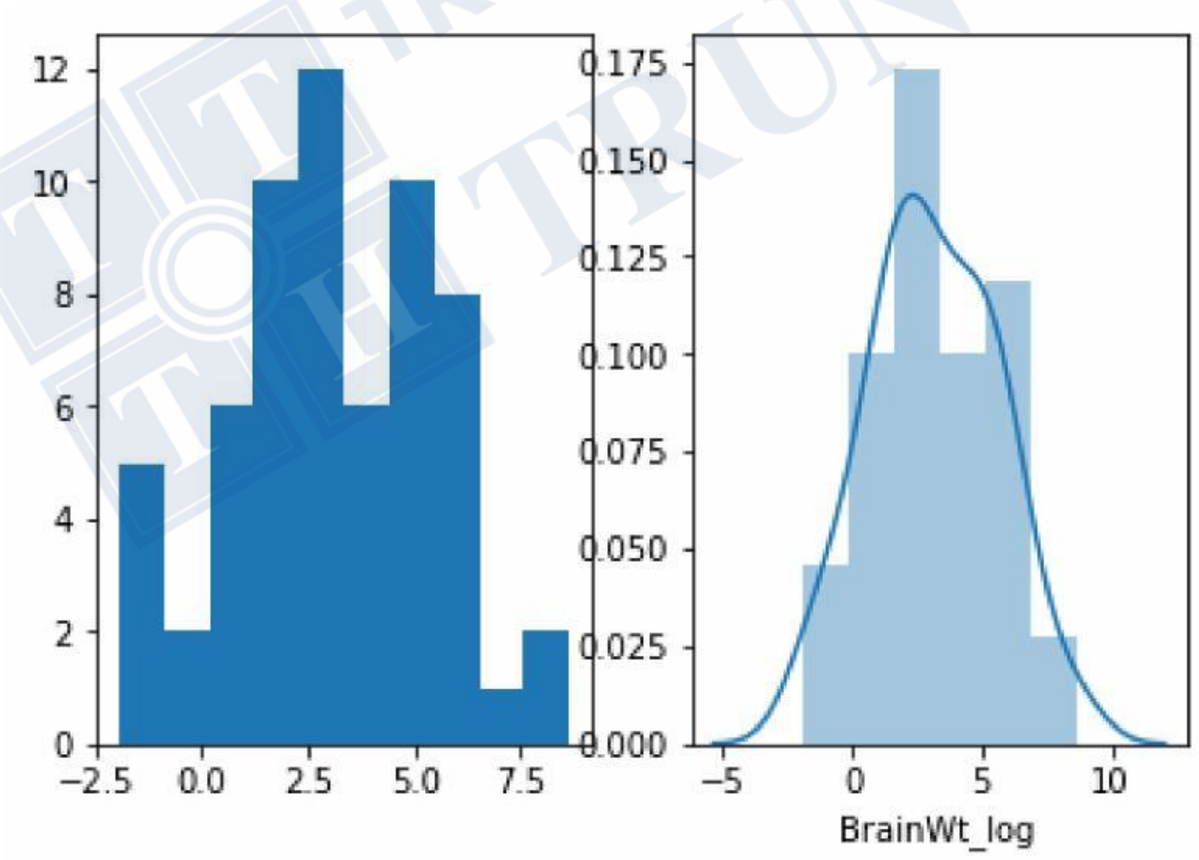
```
data.head()
```

Out[15]:

	Species	BodyWt	BrainWt	NonDreaming	Dreaming	TotalSleep	LifeSpan	C
1	Africanelephant	6654.000	5712.0	NaN	NaN	3.3	38.6	
2	Africangiantpouchedrat	1.000	6.6	6.3	2.0	8.3	4.5	
3	ArcticFox	3.385	44.5	NaN	NaN	12.5	14.0	
4	Arcticgroundsquirrel	0.920	5.7	NaN	NaN	16.5	NaN	
5	Asianelephant	2547.000	4603.0	2.1	1.8	3.9	69.0	

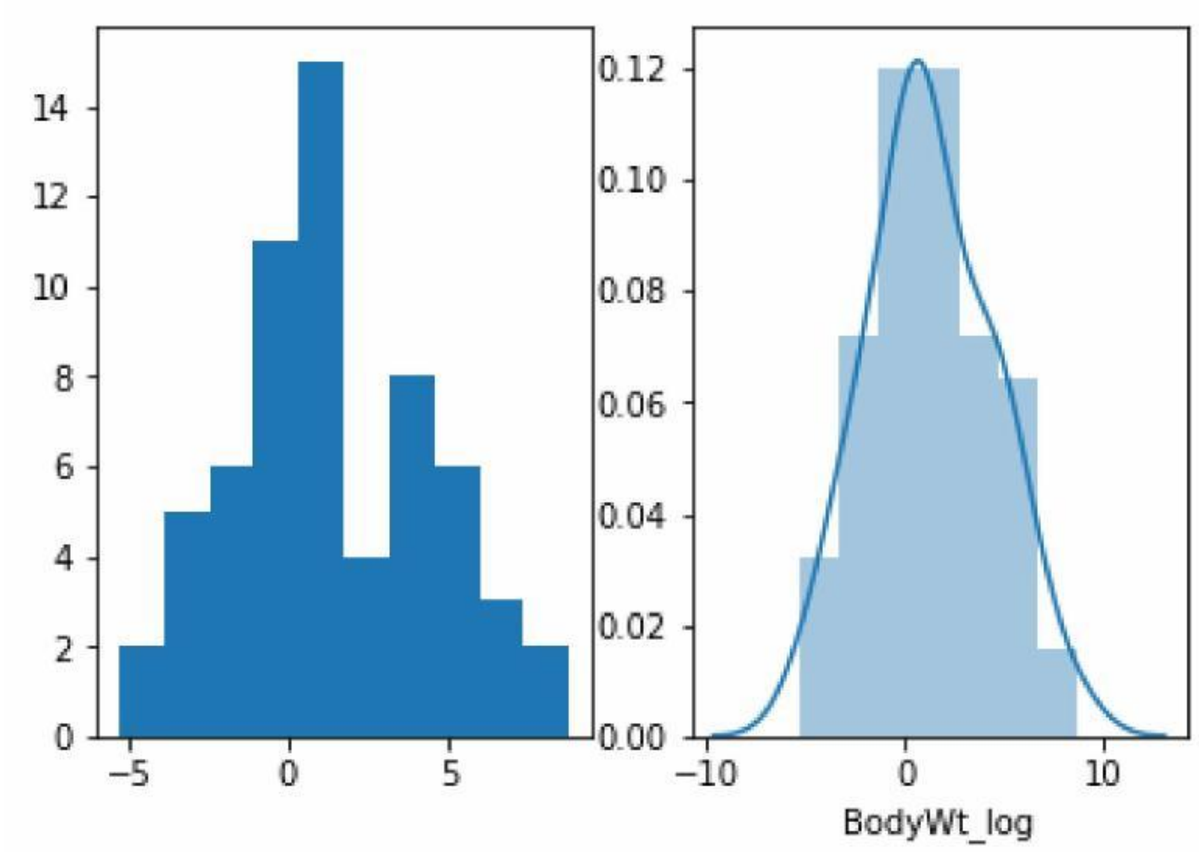
In [16]:

```
plt.subplot(1,2,1)
plt.hist(data.BrainWt_log)
plt.subplot(1,2,2)
sns.distplot(data.BrainWt_log)
plt.show()
```



In [17]:

```
plt.subplot(1,2,1)
plt.hist(data.BodyWt_log)
plt.subplot(1,2,2)
sns.distplot(data.BodyWt_log)
plt.show()
```

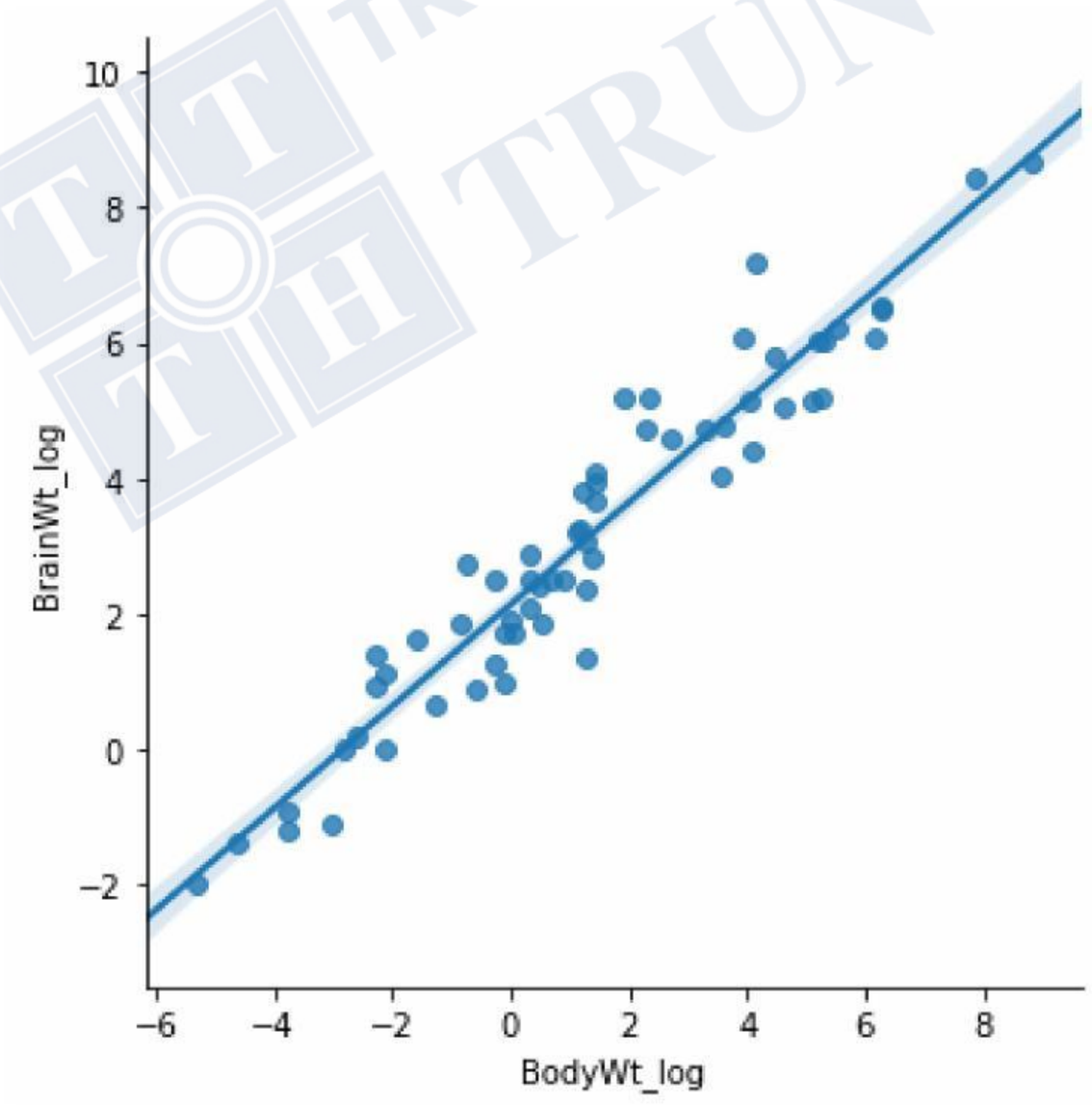


In [18]:

```
# Xem xét mối quan hệ
sns.lmplot(data=data, x='BodyWt_log', y='BrainWt_log')
```

Out[18]:

```
<seaborn.axisgrid.FacetGrid at 0x215334ef470>
```



In [ ]:

```

```