In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
```

In [10]:
```
df = pd.read_csv('marvel_box_office.csv',encoding = 'iso-8859-1')
```

In [12]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 66 entries, 0 to 65
Data columns (total 21 columns):
 #   Column                                Non-Null Count  Dtype  
---  ------                                --------------  -----  
 0   Movie                                 66 non-null     object 
 1   Release Date                          66 non-null     object 
 2   Release Month                         66 non-null     object 
 3   Release Day                           66 non-null     int64  
 4   Release Year                          66 non-null     int64  
 5   Ownership                             66 non-null     object 
 6   Domestic Box Office                   66 non-null     int64  
 7   Inflation Adjusted Domestic           66 non-null     int64  
 8   International Box Office              66 non-null     int64  
 9   Inflation Adjusted International(Dalah)  66 non-null  float64
 10  Worldwide Box Office                  66 non-null     int64  
 11  Inflation Adjusted Worldwide          66 non-null     float64
 12  Opening Weekend                       66 non-null     int64  
 13  Budget                                66 non-null     int64  
 14  IMDb Score                            66 non-null     float64
 15  Meta Score                            66 non-null     float64
 16  Tomatometer                           66 non-null     int64  
 17  Rotten Tomato Audience Score          66 non-null     int64  
 18  Run Time In Minutes                   66 non-null     int64  
 19  Phase                                 33 non-null     object 
 20  Director                              66 non-null     object 
dtypes: float64(4), int64(11), object(6)
memory usage: 11.0+ KB
```

In [13]:
```
df.head()
```

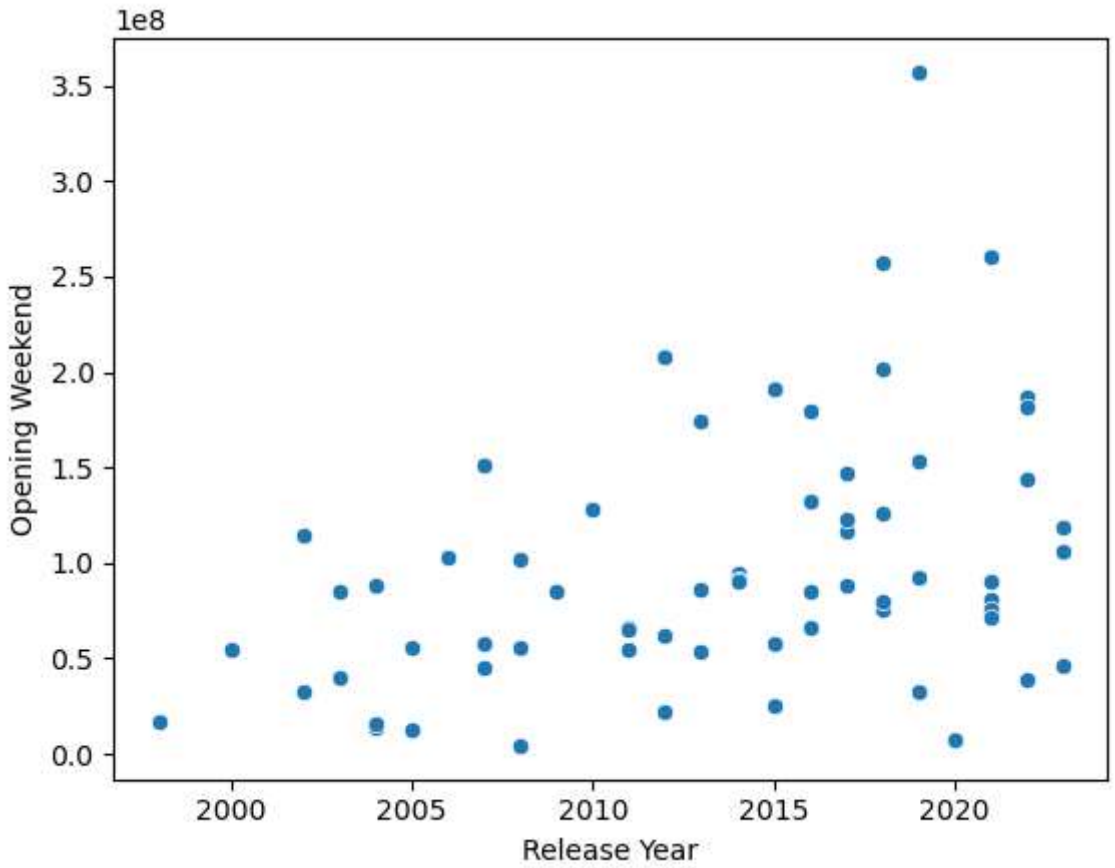
Out[13]:

	Movie	Release Date	Release Month	Release Day	Release Year	Ownership	Domestic Box Office	Inflation Adjusted Domestic	Internation Box Offi
0	Iron Man	5/2/2008	May	2	2008	Marvel Studios	318604126	467231126	2665674
1	The Incredible Hulk	6/13/2008	June	13	2008	Marvel Studios	134806913	197704288	1307669
2	Iron Man 2	5/7/2010	May	7	2010	Marvel Studios	312433331	416973763	3087230
3	Thor	5/6/2011	May	6	2011	Marvel Studios	181030624	240384926	2682959
4	Captain America: The First Avenger	7/22/2011	July	22	2011	Marvel Studios	176654505	234574020	1939152

5 rows × 21 columns

In [14]:
```
sns.scatterplot(data = df , x = 'Release Year', y = 'Opening Weekend')
```

Out[14]:
```
<Axes: xlabel='Release Year', ylabel='Opening Weekend'>
```



In [17]:
```
df2 = df[ ['Release Year', 'Opening Weekend']].dropna()
df2.head()
```

Out[17]:

	Release Year	Opening Weekend
0	2008	102118668
1	2008	55414050
2	2010	128122480
3	2011	65723338
4	2011	65058524

In [18]:
```
from sklearn.cluster import KMeans
```

In [25]:
```
model = KMeans(n_clusters=3,random_state=0)
model.fit(df2)
```

```
c:\Users\User\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1412: Fu
tureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  super()._check_params_vs_input(X, default_n_init=10)
c:\Users\User\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1436: Us
erWarning: KMeans is known to have a memory leak on Windows with MKL, when th
ere are less chunks than available threads. You can avoid it by setting the e
nvironment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

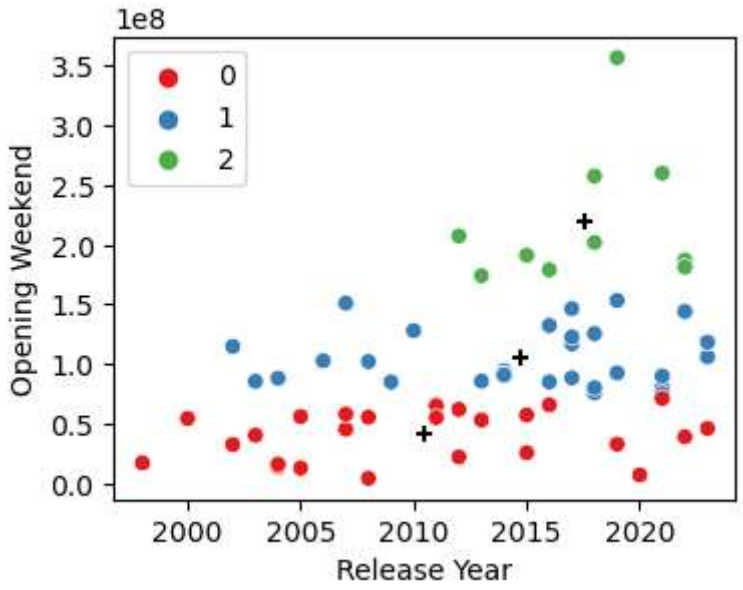
Out[25]:
```
KMeans(n_clusters=3, random_state=0)
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

In [26]:
```
model.cluster_centers_
```

Out[26]:
```
array([[2.01046154e+03, 4.13211263e+07],
       [2.01470000e+03, 1.04983068e+08],
       [2.01760000e+03, 2.19771001e+08]])
```

In [27]:
```
plt.figure(figsize = [4,3])
sns.scatterplot(data = df2 , x = 'Release Year' , y = 'Opening Weekend', hue =
plt.scatter(model.cluster_centers_[:,0], model.cluster_centers_[:,1], color = '
```

Out[27]:
```
<matplotlib.collections.PathCollection at 0x1f1facaec90>
```



In [28]:
```
model.labels_
```

Out[28]:
```
array([1, 0, 1, 0, 0, 2, 2, 1, 1, 1, 2, 0, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2,
       1, 1, 1, 0, 2, 2, 1, 2, 1, 1, 0, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 1,
       1, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])
```

In [29]:
```
model.predict([[200,100],[350,150]])
```

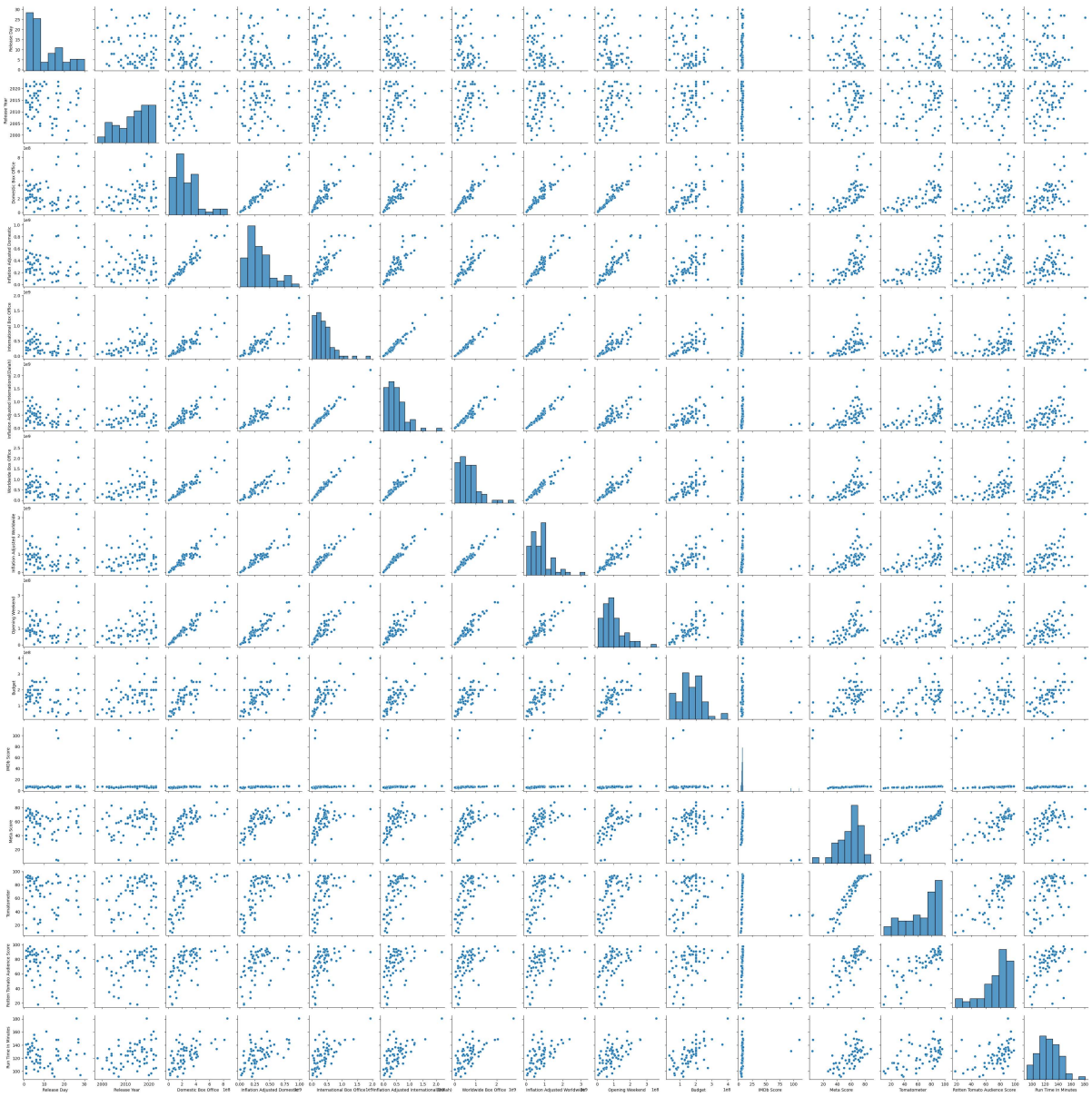
```
c:\Users\User\anaconda3\Lib\site-packages\sklearn\base.py:464: UserWarning: X
does not have valid feature names, but KMeans was fitted with feature names
  warnings.warn(
```

Out[29]:
```
array([0, 0])
```

In [30]:
```
sns.pairplot(df)
```

```
c:\Users\User\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarnin
g: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[30]:
```
<seaborn.axisgrid.PairGrid at 0x1f1f9bb6d90>
```



In [ ]:
```

```