In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
```

In [3]:
```
df = pd.read_csv('marvel_box_office.csv',encoding = 'iso-8859-1')
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  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 [4]:
```
df.head()
```

Out[4]:

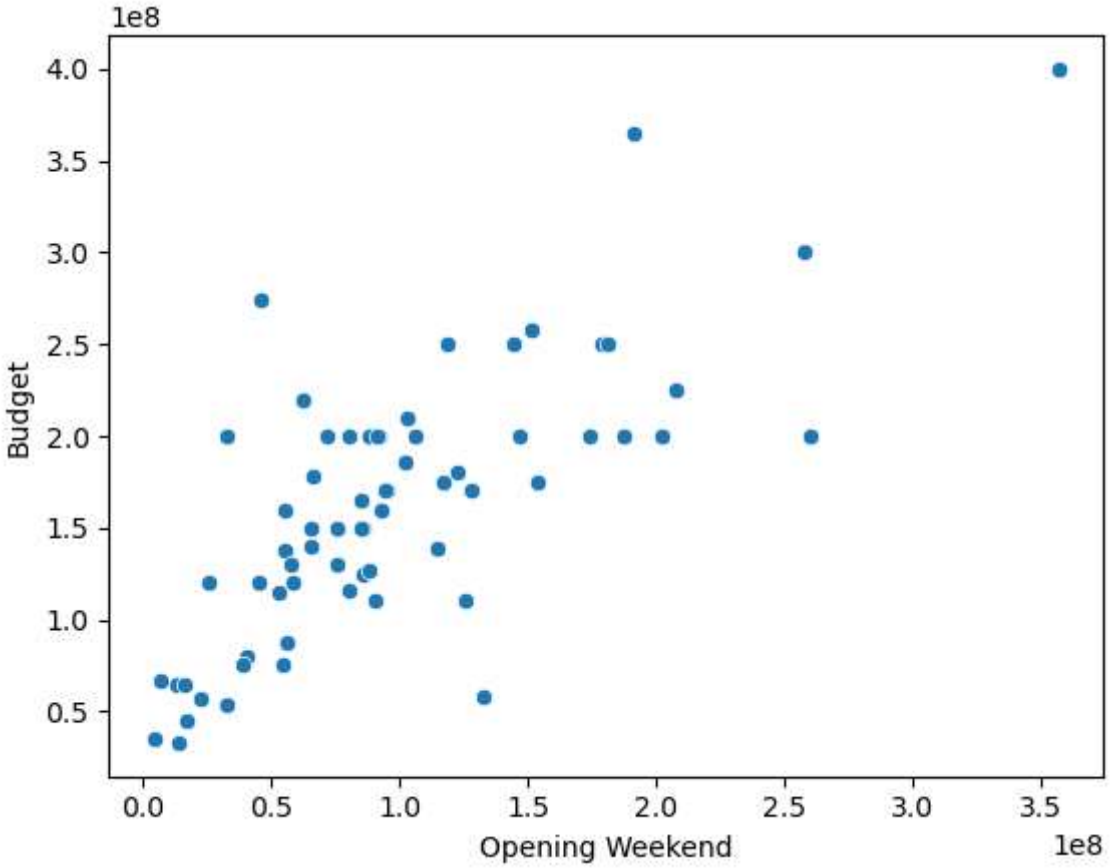
	Movie	Release Date	Release Month	Release Day	Release Year	Ownership	Domestic Box Office	Inflation Adjusted Domestic	Internatio Box Offi
0	Iron Man	5/2/2008	May	2	2008	Marvel Studios	318604126	467231126	2665674
1	The Incredible Hulk	6/13/2008	June	13	2008	Marvel Studios	134806913	197704288	1307669
2	Iron 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 [5]:
```
sns.scatterplot(data=df,x='Opening Weekend',y='Budget', palette='Set1')
```

```
C:\Users\walo1\AppData\Local\Temp\ipykernel_13196\600821124.py:1: UserWarnin
g: Ignoring `palette` because no `hue` variable has been assigned.
  sns.scatterplot(data=df,x='Opening Weekend',y='Budget', palette='Set1')
```

Out[5]:
```
<Axes: xlabel='Opening Weekend', ylabel='Budget'>
```



In [6]:
```
model = KMeans(n_clusters=3, random_state=0)
df2 = df[['Opening Weekend','Budget']].dropna()
model.fit(df2)
```

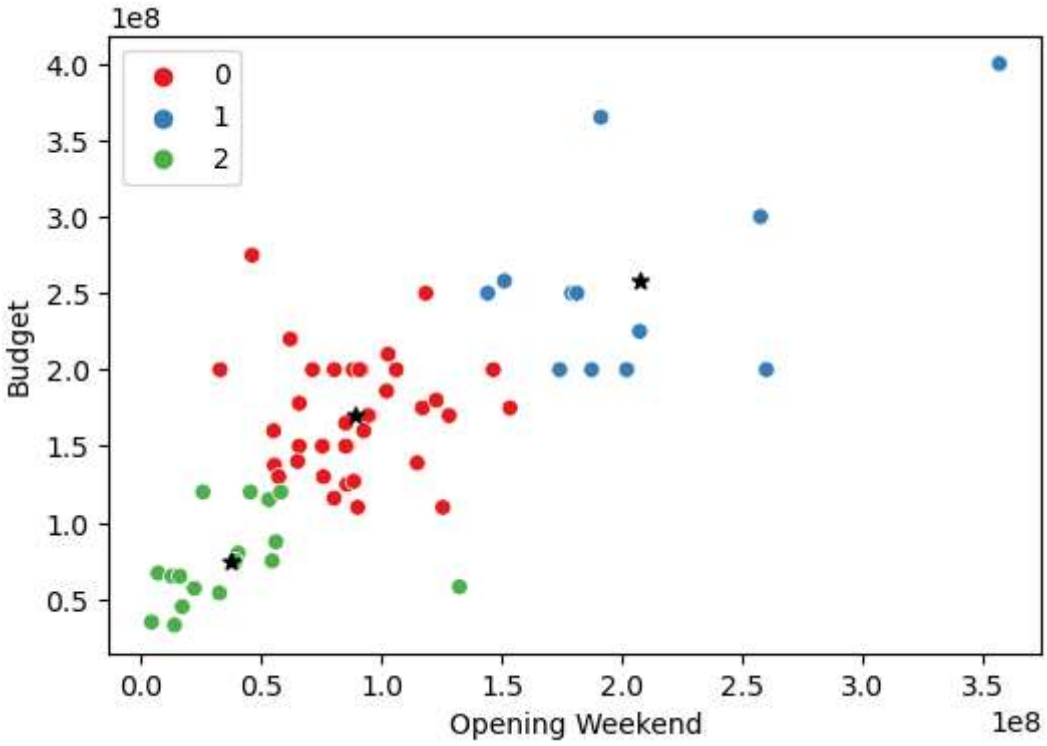
```
c:\Users\walo1\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1412: F
utureWarning: The default value of `n_init` will change from 10 to 'auto' in
1.4. Set the value of `n_init` explicitly to suppress the warning
  super()._check_params_vs_input(X, default_n_init=10)
c:\Users\walo1\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1436: U
serWarning: KMeans is known to have a memory leak on Windows with MKL, when t
here are less chunks than available threads. You can avoid it by setting the
environment variable OMP_NUM_THREADS=1.
  warnings.warn(
```

Out[6]:
```
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 [7]:
```
plt.figure(figsize=[6,4])
sns.scatterplot(data=df2,x='Opening Weekend',y='Budget', hue=model.labels_,pal
plt.scatter(model.cluster_centers_[:,0],model.cluster_centers_[:,1],color='k',
```

Out[7]:
```
<matplotlib.collections.PathCollection at 0x176daf92d10>
```



In [8]:
```
model.predict([ [ 200,100 ] ,[ 350 , 150 ]])
```

```
c:\Users\walo1\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[8]:
```
array([2, 2])
```

In [9]:
```
sns.pairplot(df)
```

```
c:\Users\walo1\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarni
ng: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```

Out[9]: <seaborn.axisgrid.PairGrid at 0x176db6e5290>