In [1]:
```
import pandas as pd
```

In [2]:
```
movies = pd.read_csv(r'C:\Users\rohit\OneDrive\Desktop\FSDS & GEN-AI 25TH-NOV-02
```

In [3]:
```
movies
```

Out[3]:

	Film	Genre	Rotten Tomatoes Ratings %	Audience Ratings %	Budget (million $)	Year of release
0	(500) Days of Summer	Comedy	87	81	8	2009
1	10,000 B.C.	Adventure	9	44	105	2008
2	12 Rounds	Action	30	52	20	2009
3	127 Hours	Adventure	93	84	18	2010
4	17 Again	Comedy	55	70	20	2009
...	...	...	...	...	...	...
554	Your Highness	Comedy	26	36	50	2011
555	Youth in Revolt	Comedy	68	52	18	2009
556	Zodiac	Thriller	89	73	65	2007
557	Zombieland	Action	90	87	24	2009
558	Zookeeper	Comedy	14	42	80	2011

559 rows × 6 columns

In [4]:
```
len(movies)
```

Out[4]:
```
559
```

In [5]:
```
movies.shape
```

Out[5]:
```
(559, 6)
```

In [6]:
```
movies.columns
```

Out[6]:
```
Index(['Film', 'Genre', 'Rotten Tomatoes Ratings %', 'Audience Ratings %',
       'Budget (million $)', 'Year of release'],
      dtype='object')
```

In [7]:
```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 559 entries, 0 to 558
Data columns (total 6 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   Film                       559 non-null    object
 1   Genre                      559 non-null    object
 2   Rotten Tomatoes Ratings %  559 non-null    int64 
 3   Audience Ratings %         559 non-null    int64 
 4   Budget (million $)         559 non-null    int64 
 5   Year of release            559 non-null    int64 
dtypes: int64(4), object(2)
memory usage: 26.3+ KB
```

In [8]:
```
movies.head()
```

Out[8]:

	Film	Genre	Rotten Tomatoes Ratings %	Audience Ratings %	Budget (million $)	Year of release
0	(500) Days of Summer	Comedy	87	81	8	2009
1	10,000 B.C.	Adventure	9	44	105	2008
2	12 Rounds	Action	30	52	20	2009
3	127 Hours	Adventure	93	84	18	2010
4	17 Again	Comedy	55	70	20	2009

In [9]:
```
movies.tail()
```

Out[9]:

	Film	Genre	Rotten Tomatoes Ratings %	Audience Ratings %	Budget (million $)	Year of release
554	Your Highness	Comedy	26	36	50	2011
555	Youth in Revolt	Comedy	68	52	18	2009
556	Zodiac	Thriller	89	73	65	2007
557	Zombieland	Action	90	87	24	2009
558	Zookeeper	Comedy	14	42	80	2011

In [10]:
```
import numpy
print(numpy.__version__)
```

```
1.26.4
```

In [11]:
```
import pandas
print(pandas.__version__)
```

```
2.2.2
```

In [12]:
```
movies.columns
```

Out[12]:
```
Index(['Film', 'Genre', 'Rotten Tomatoes Ratings %', 'Audience Ratings %',
       'Budget (million $)', 'Year of release'],
      dtype='object')
```

In [13]:
```
movies.columns = ['Film', 'Genre', 'CriticRating', 'AudienceRating', 'BudgetMill
```

In [14]:
```
movies.columns
```

Out[14]:
```
Index(['Film', 'Genre', 'CriticRating', 'AudienceRating', 'BudgetMillions',
       'Year'],
      dtype='object')
```

In [15]:
```
movies.head(1)  # Removed spaces & % removed noise characters
```

Out[15]:

	Film	Genre	CriticRating	AudienceRating	BudgetMillions	Year
0	(500) Days of Summer	Comedy	87	81	8	2009

In [16]:
```
movies.shape
```

Out[16]:
```
(559, 6)
```

In [17]:
```
movies.describe()
```

Out[17]:

	CriticRating	AudienceRating	BudgetMillions	Year
count	559.000000	559.000000	559.000000	559.000000
mean	47.309481	58.744186	50.236136	2009.152057
std	26.413091	16.826887	48.731817	1.362632
min	0.000000	0.000000	0.000000	2007.000000
25%	25.000000	47.000000	20.000000	2008.000000
50%	46.000000	58.000000	35.000000	2009.000000
75%	70.000000	72.000000	65.000000	2010.000000
max	97.000000	96.000000	300.000000	2011.000000

In [18]:
```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 559 entries, 0 to 558
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   Film            559 non-null    object
 1   Genre           559 non-null    object
 2   CriticRating    559 non-null    int64 
 3   AudienceRating  559 non-null    int64 
 4   BudgetMillions  559 non-null    int64 
 5   Year            559 non-null    int64 
dtypes: int64(4), object(2)
memory usage: 26.3+ KB
```

In [19]:
```
movies.Film = movies.Film.astype('category')
```

In [20]:
```
movies.Film
```

Out[20]:
```
0      (500) Days of Summer
1                10,000 B.C.
2                 12 Rounds
3                  127 Hours
4                  17 Again
               ...
554           Your Highness
555         Youth in Revolt
556                  Zodiac
557             Zombieland
558               Zookeeper
Name: Film, Length: 559, dtype: category
Categories (559, object): ['(500) Days of Summer ', '10,000 B.C.', '12 Rounds
', '127 Hours', ..., 'Youth in Revolt', 'Zodiac', 'Zombieland ', 'Zookeeper']
```

In [21]:
```
movies.Genre = movies.Genre.astype('category')
movies.Year = movies.Year.astype('category')
```

In [22]:
```
movies.Genre
```

Out[22]:
```
0         Comedy
1      Adventure
2         Action
3      Adventure
4         Comedy
          ...
554       Comedy
555       Comedy
556     Thriller
557       Action
558       Comedy
Name: Genre, Length: 559, dtype: category
Categories (7, object): ['Action', 'Adventure', 'Comedy', 'Drama', 'Horror', 'R
omance', 'Thriller']
```

In [23]:
```
movies.Year
```

Out[23]:
```
0      2009
1      2008
2      2009
3      2010
4      2009
       ...
554    2011
555    2009
556    2007
557    2009
558    2011
Name: Year, Length: 559, dtype: category
Categories (5, int64): [2007, 2008, 2009, 2010, 2011]
```

In [24]:
```
movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 559 entries, 0 to 558
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Film            559 non-null    category
 1   Genre           559 non-null    category
 2   CriticRating    559 non-null    int64
 3   AudienceRating  559 non-null    int64
 4   BudgetMillions  559 non-null    int64
 5   Year            559 non-null    category
dtypes: category(3), int64(3)
memory usage: 36.5 KB
```

In [25]:
```
movies.describe()
```

Out[25]:

	CriticRating	AudienceRating	BudgetMillions
count	559.000000	559.000000	559.000000
mean	47.309481	58.744186	50.236136
std	26.413091	16.826887	48.731817
min	0.000000	0.000000	0.000000
25%	25.000000	47.000000	20.000000
50%	46.000000	58.000000	35.000000
75%	70.000000	72.000000	65.000000
max	97.000000	96.000000	300.000000

In [26]:
```
from matplotlib import pyplot as plt
import seaborn as sns
```

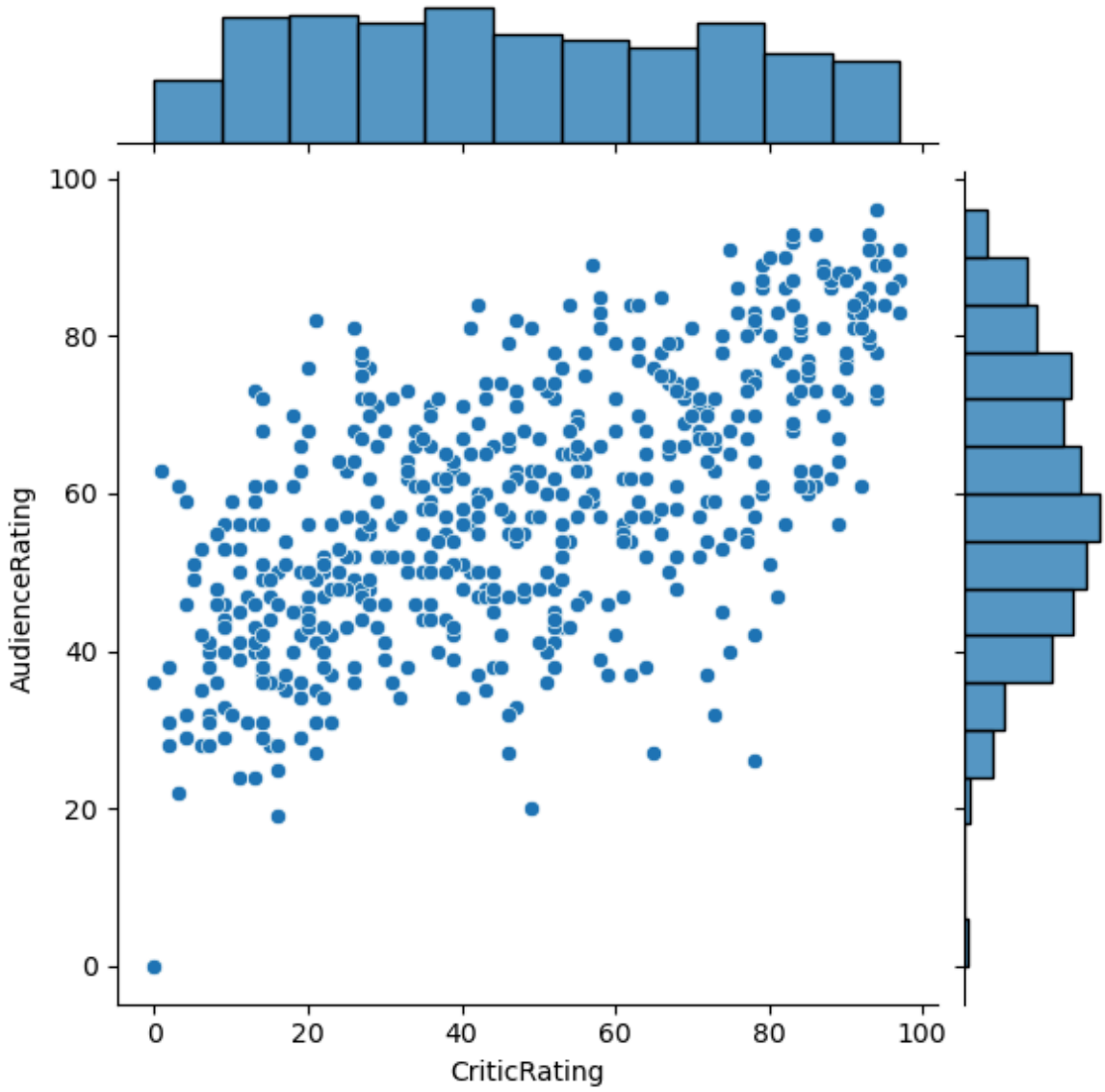
In [27]:
```
%matplotlib inline
```

In [28]:
```
import warnings
warnings.filterwarnings('ignore')
```

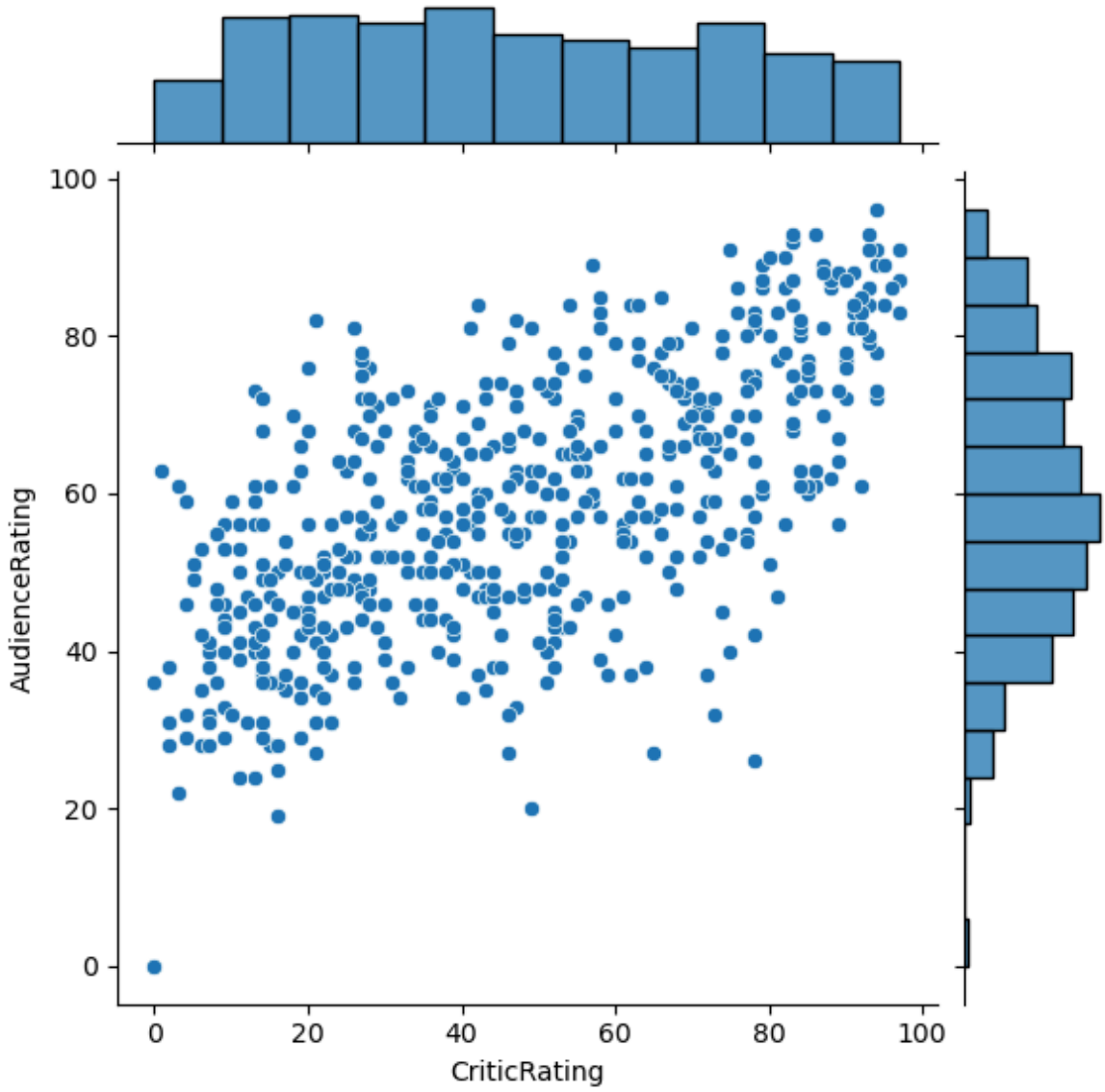
In [29]:
```
j = plt.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating')
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
Cell In[29], line 1
----> 1 j = plt.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRatin
g')

AttributeError: module 'matplotlib.pyplot' has no attribute 'jointplot'
```

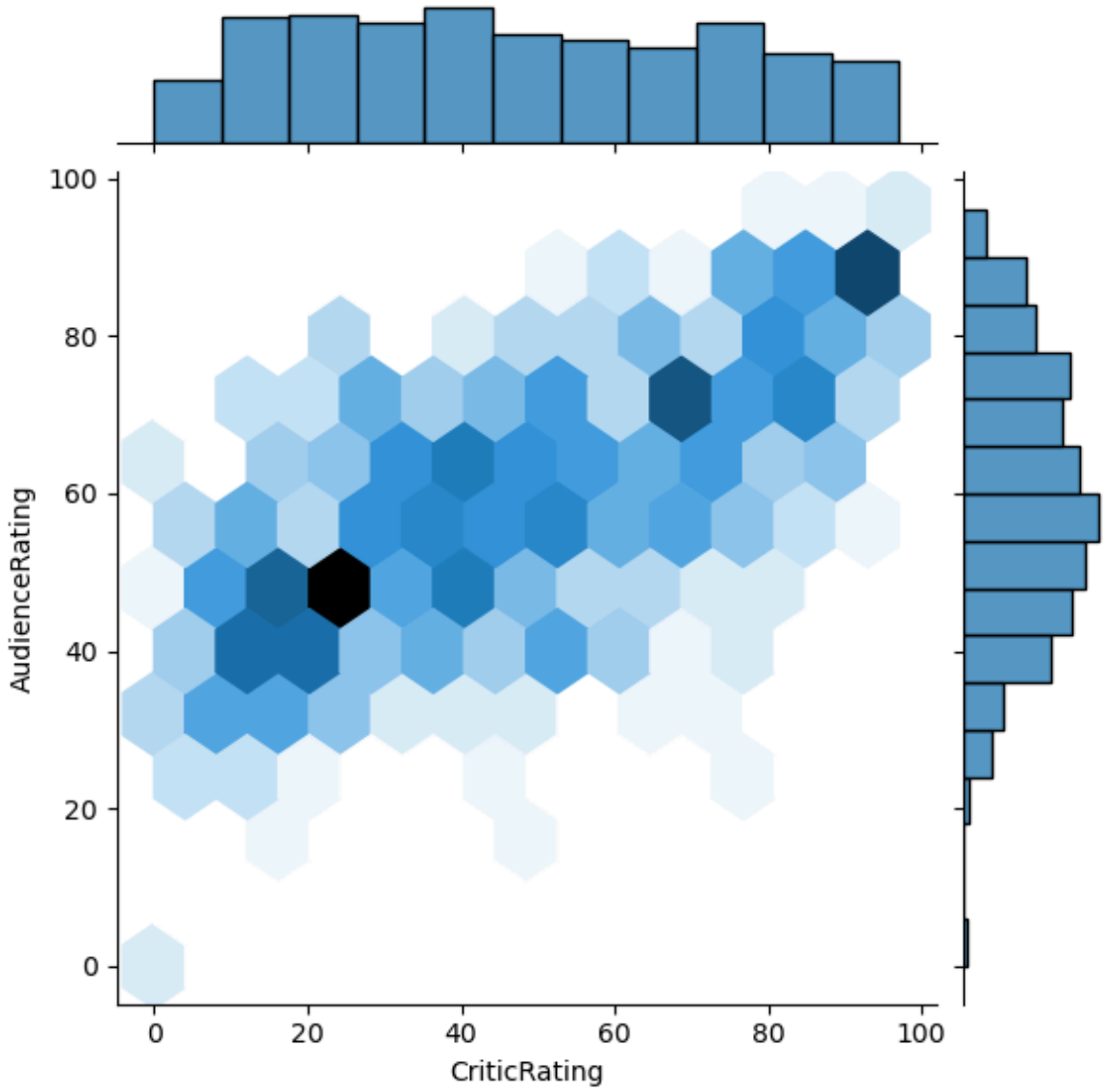
In [30]:
```
j = sns.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating')
plt.show(j)
```



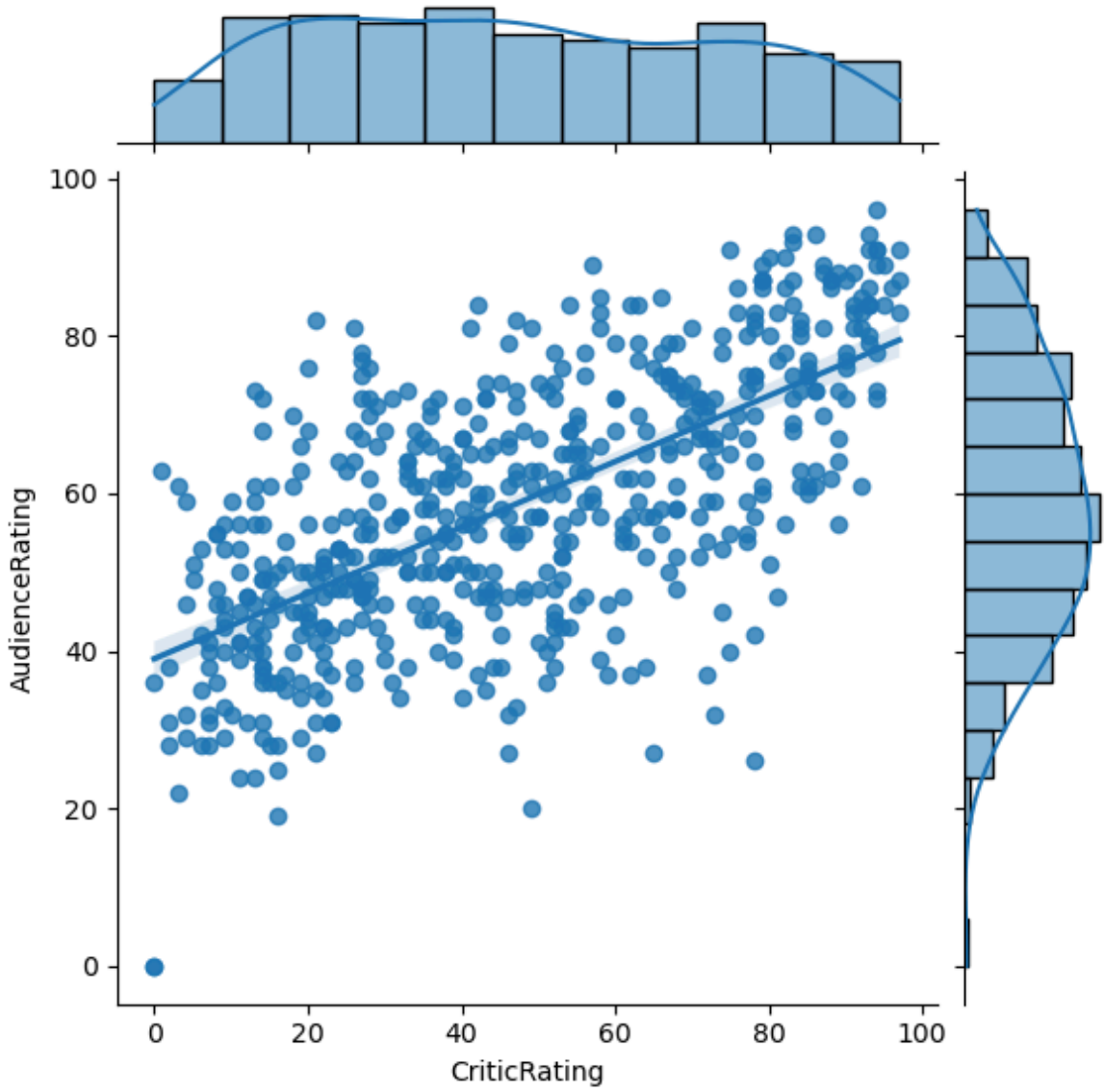
In [31]:
```
j = sns.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating', kind
plt.show(j)
```



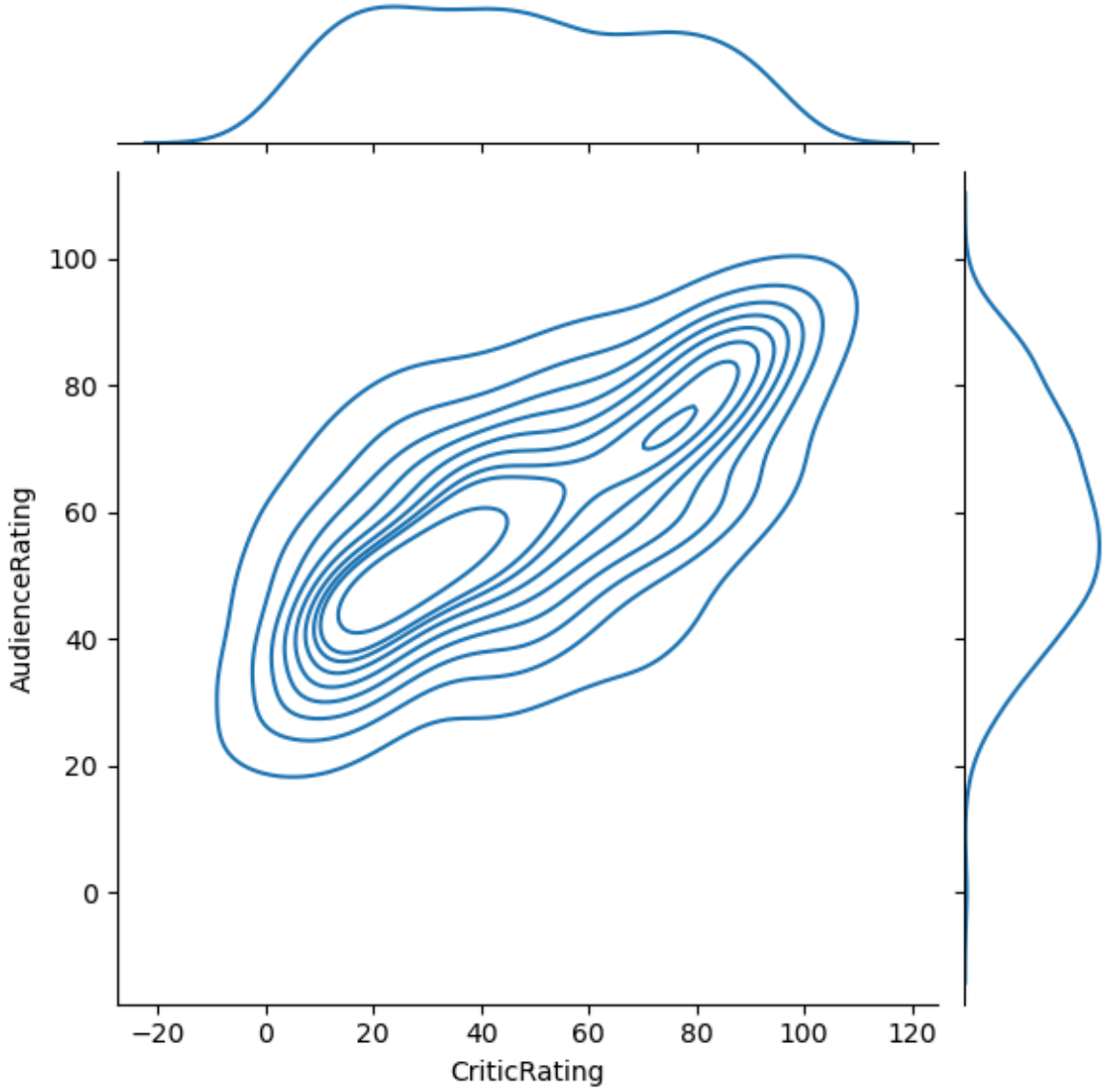
In [34]:
```
j = sns.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating', kind
plt.show(j)
```



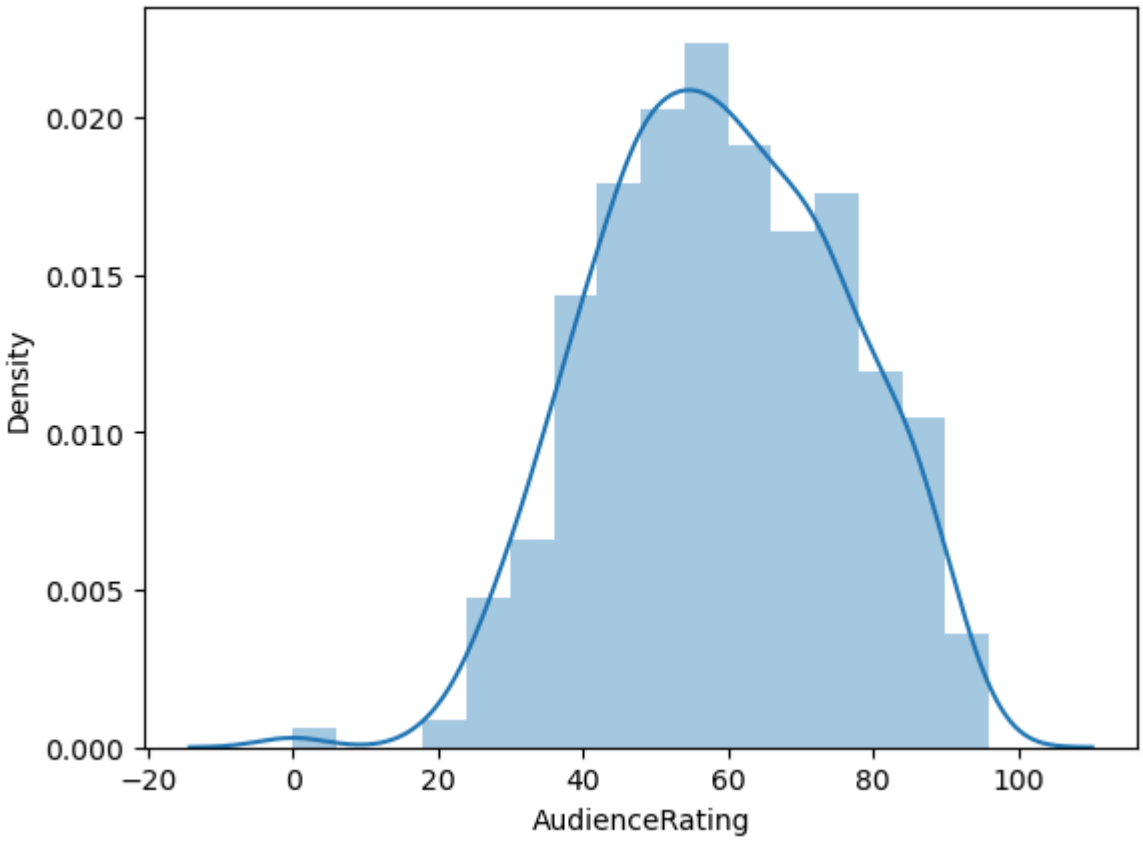
In [36]:
```
j = sns.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating', kind=
plt.show(j)
```



In [38]:
```
j = sns.jointplot(data = movies, x = 'CriticRating', y = 'AudienceRating', kind
plt.show(j)
```

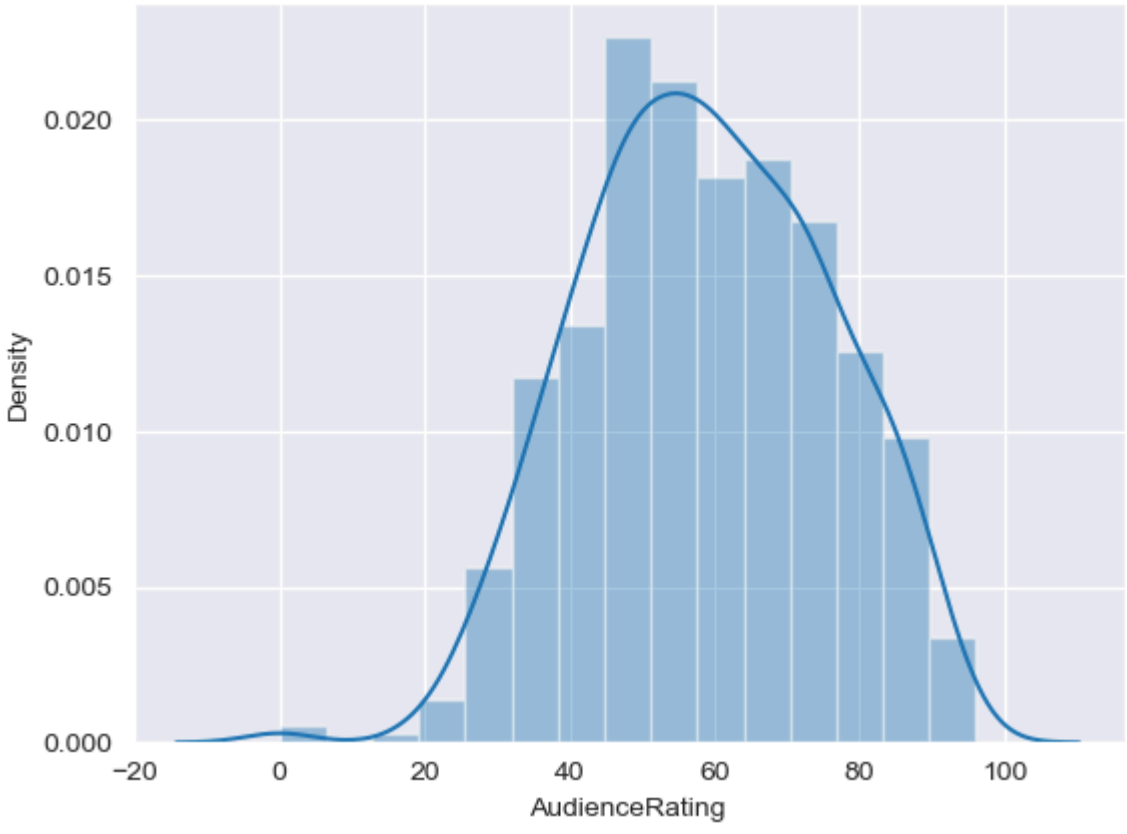


In [40]:
```
# Histograms

m1 = sns.distplot(movies.AudienceRating)
plt.show(m1)
```



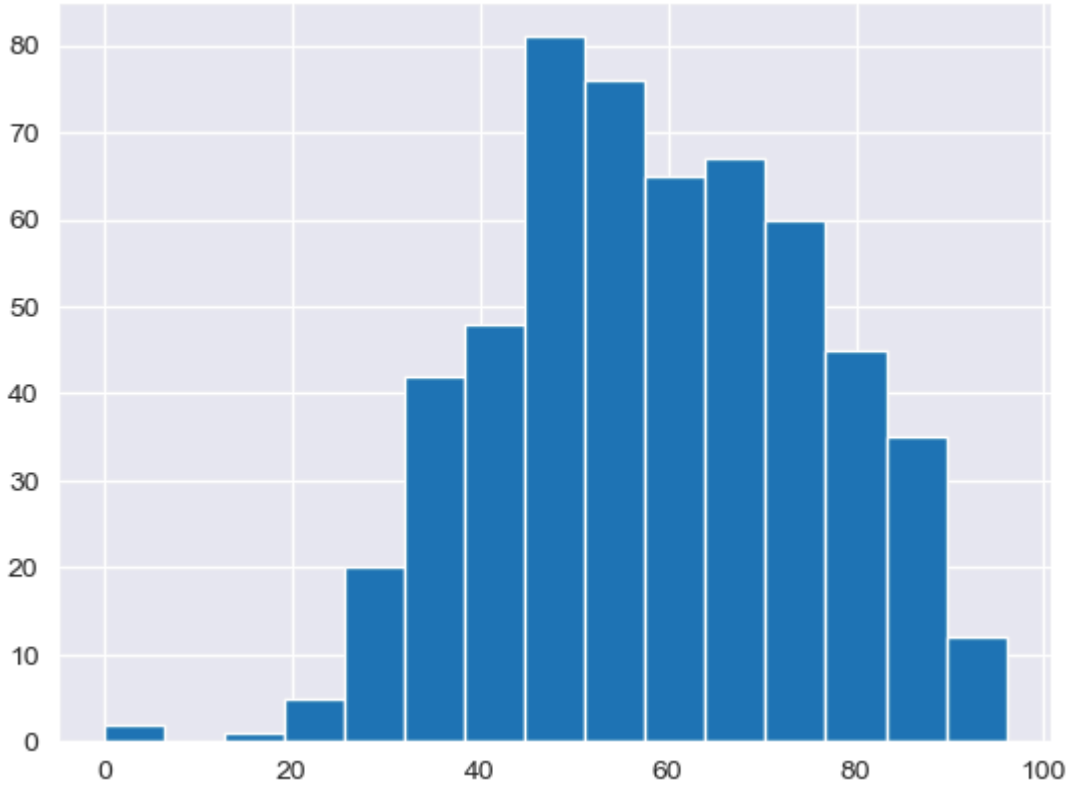
In [42]:
```
sns.set_style('darkgrid')
```

In [44]:
```
m2 = sns.distplot(movies.AudienceRating, bins = 15)
plt.show(m2)
```

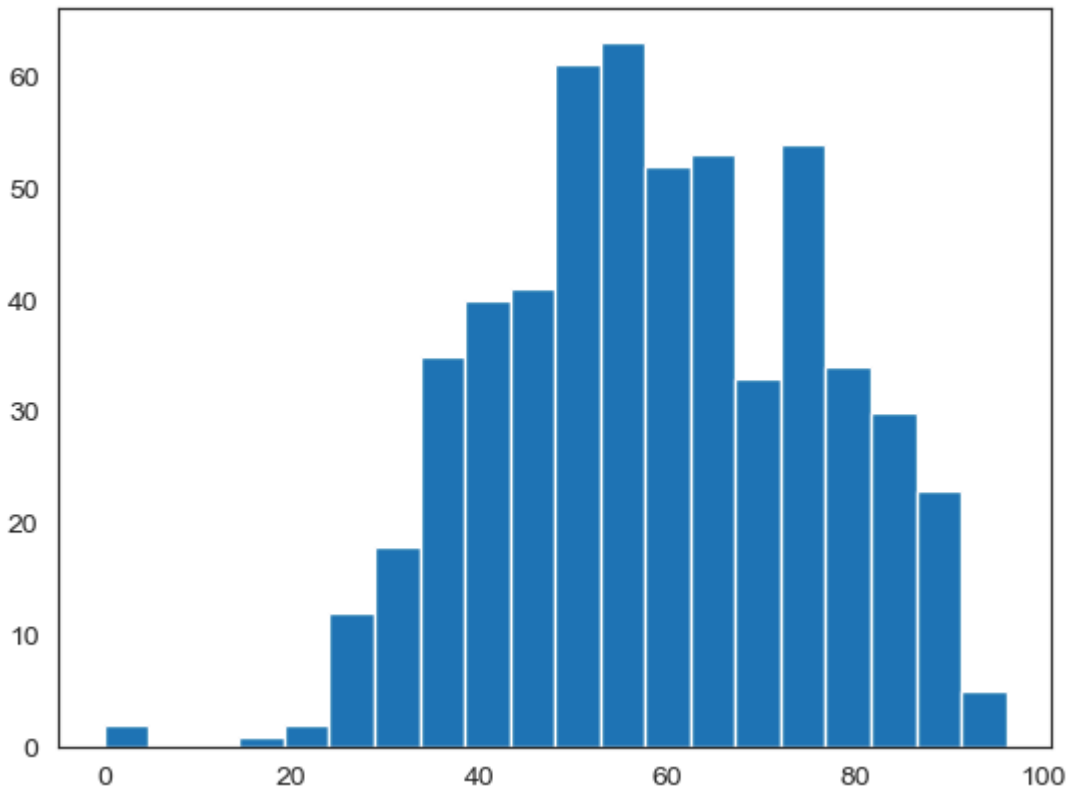


In [46]:
```
sns.set_style('darkgrid')
```

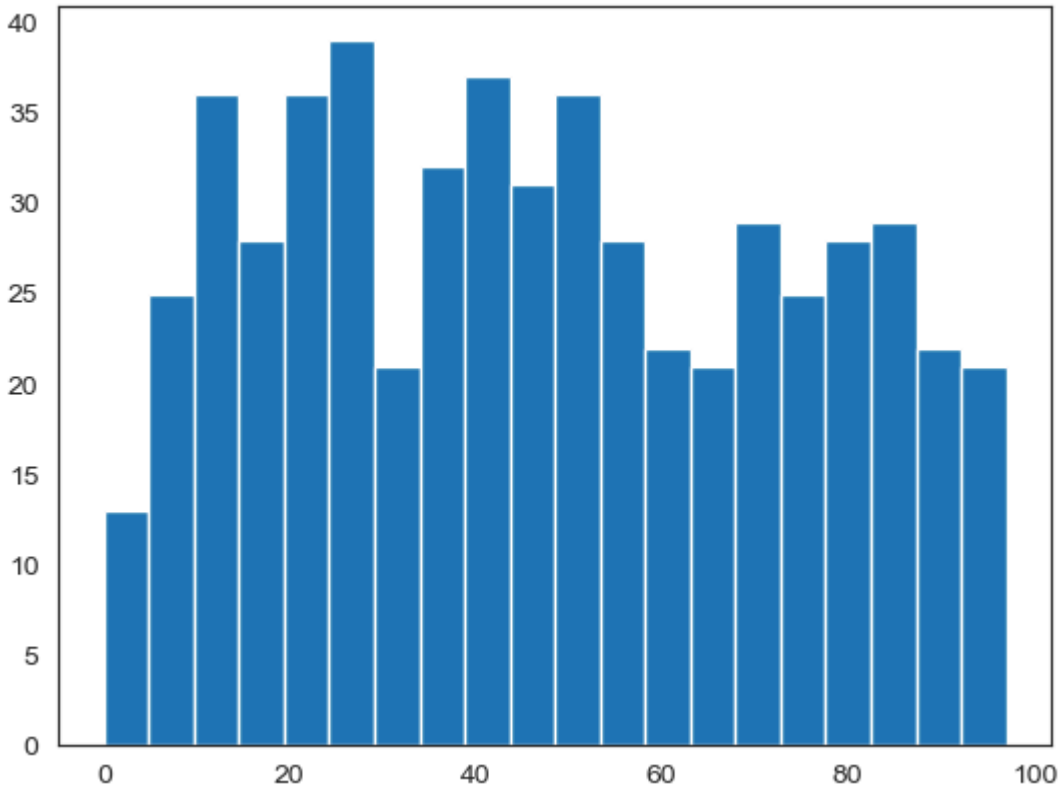
In [48]:
```
n1 = plt.hist(movies.AudienceRating, bins=15)
plt.show(n1)
```



In [50]:
```
sns.set_style('white')  # normal distribution & called as bell curve
n1 = plt.hist(movies.AudienceRating, bins=20)
plt.show(n1)
```

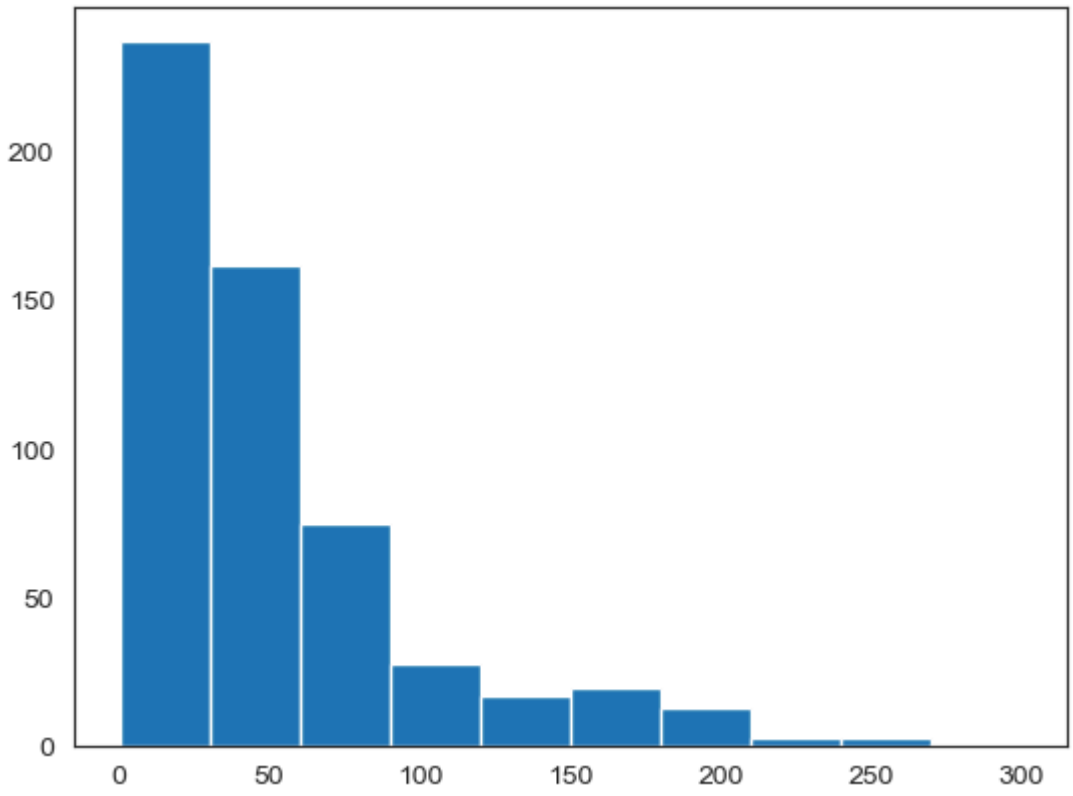


In [52]:
```
n2 = plt.hist(movies.CriticRating, bins=20)  # uniform distribution
plt.show(n2)
```

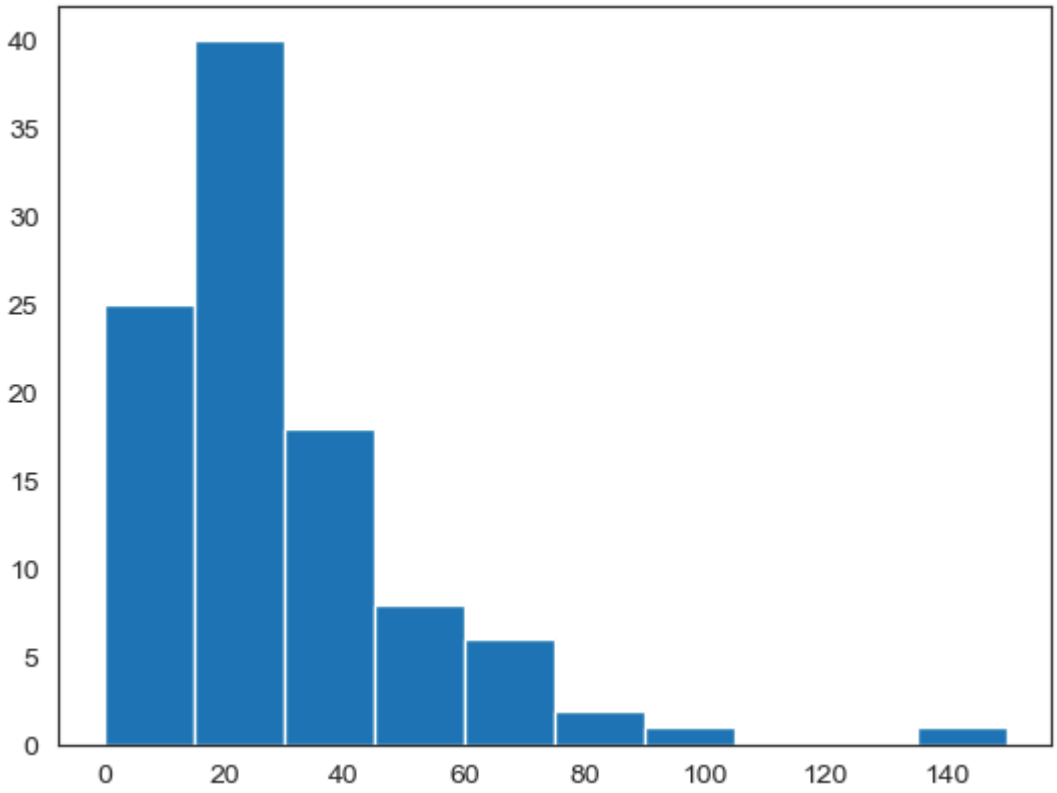


In [54]:
```
# << chat--2

# creating stacked histograms & this is tough to understand
```

In [56]:
```
# h1 = pt.hist(movies.BudgetMillions)

plt.hist(movies.BudgetMillions)
plt.show()
```



In [58]:
```
plt.hist(movies[movies.Genre == 'Drama'].BudgetMillions)
plt.show()
```



In [60]:
```
movies.head()
```

Out[60]:

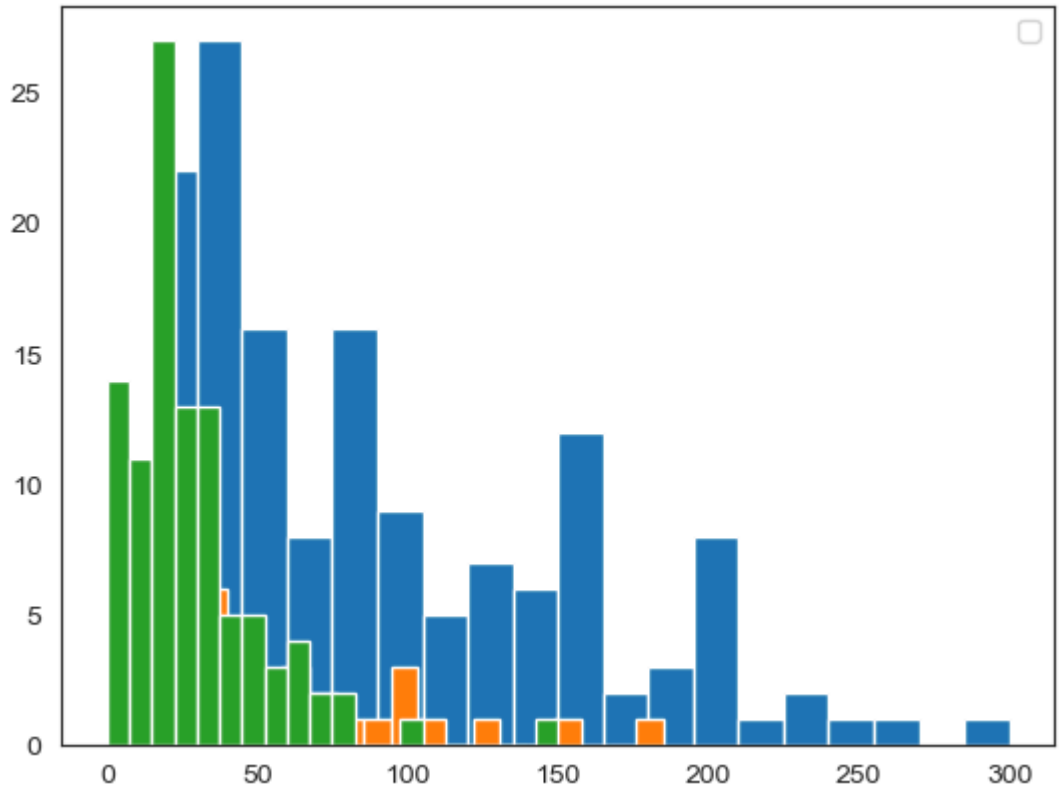
	Film	Genre	CriticRating	AudienceRating	BudgetMillions	Year
0	(500) Days of Summer	Comedy	87	81	8	2009
1	10,000 B.C.	Adventure	9	44	105	2008
2	12 Rounds	Action	30	52	20	2009
3	127 Hours	Adventure	93	84	18	2010
4	17 Again	Comedy	55	70	20	2009

In [62]:
```
movies.Genre.unique()
```

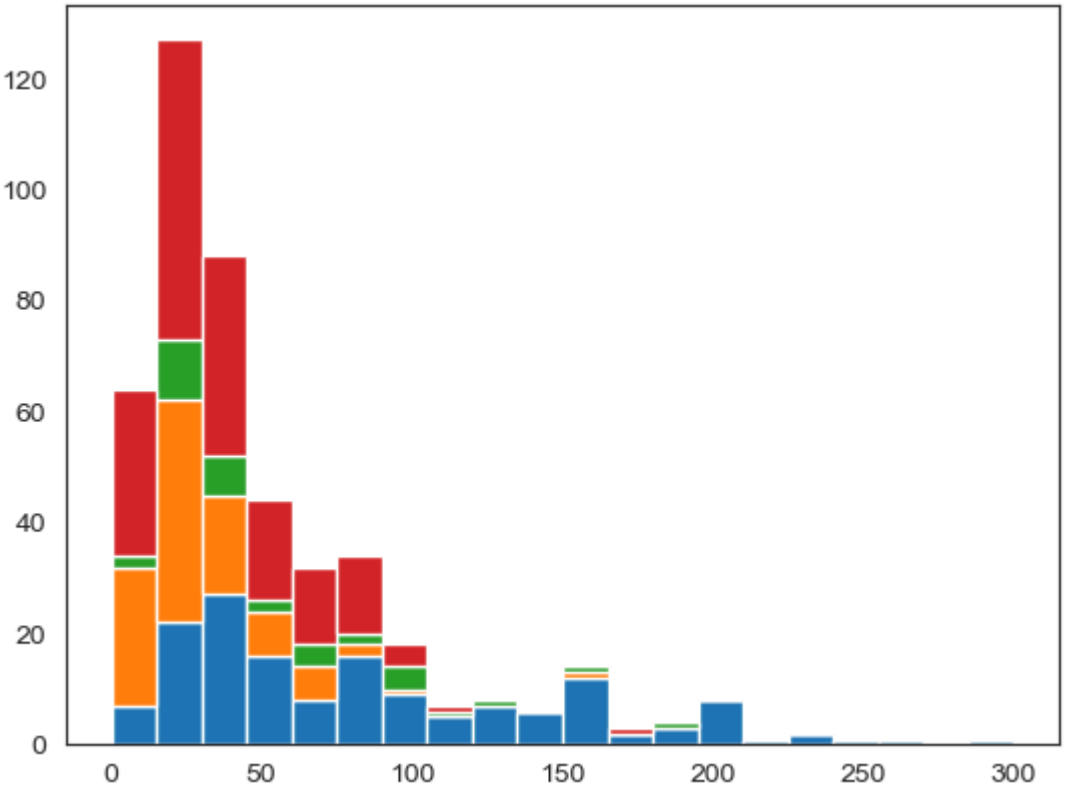
Out[62]:
```
['Comedy', 'Adventure', 'Action', 'Horror', 'Drama', 'Romance', 'Thriller']
Categories (7, object): ['Action', 'Adventure', 'Comedy', 'Drama', 'Horror', 'R
omance', 'Thriller']
```

In [64]:
```
# Below plots are stacked histogram because overlaped
```

In [70]:
```
plt.hist(movies[movies.Genre == 'Action'].BudgetMillions, bins=20)
plt.hist(movies[movies.Genre == 'Thriller'].BudgetMillions, bins=20)
plt.hist(movies[movies.Genre == 'Drama'].BudgetMillions, bins=20)
plt.legend()
plt.show()
```



In [76]:
```
plt.hist([movies[movies.Genre == 'Action'].BudgetMillions,\
          movies[movies.Genre == 'Drama'].BudgetMillions,\
          movies[movies.Genre == 'Thriller'].BudgetMillions,\
          movies[movies.Genre == 'Comedy'].BudgetMillions],
         bins = 20, stacked = True)
plt.show()
```

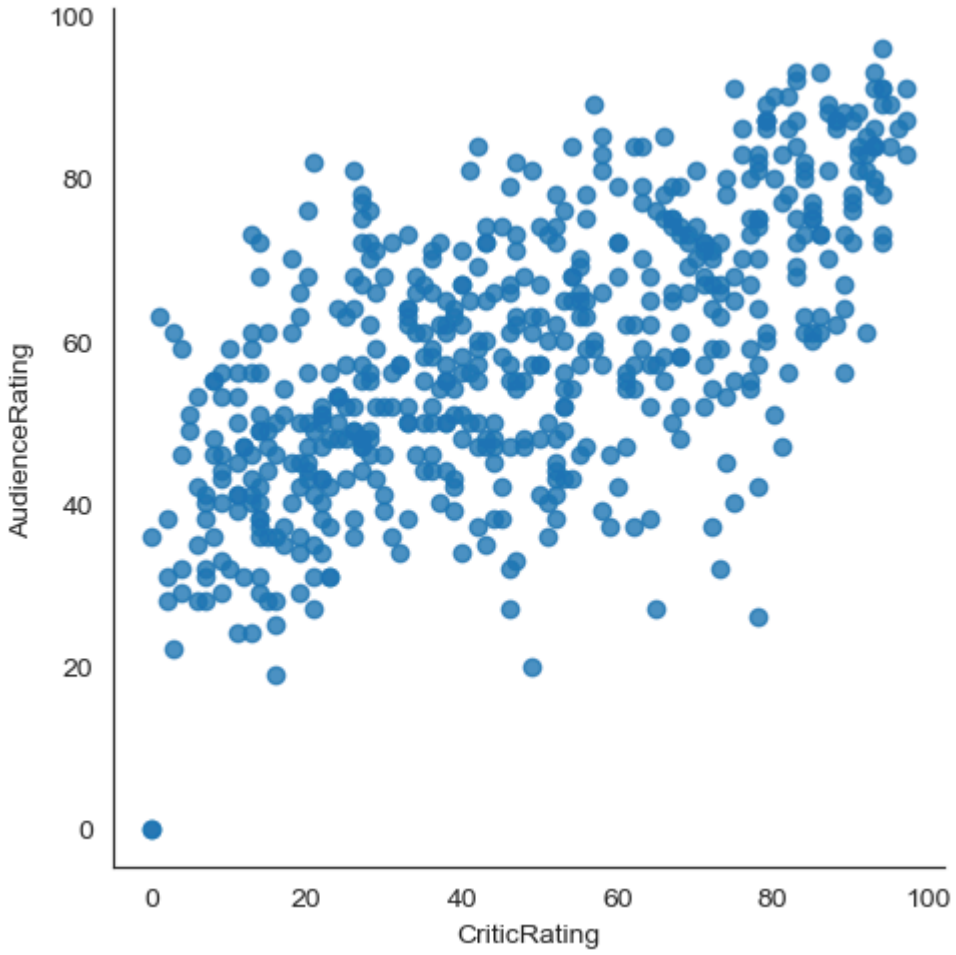


In [78]:
```
# if you have 100 categories you cannot copy & paste all the things
```

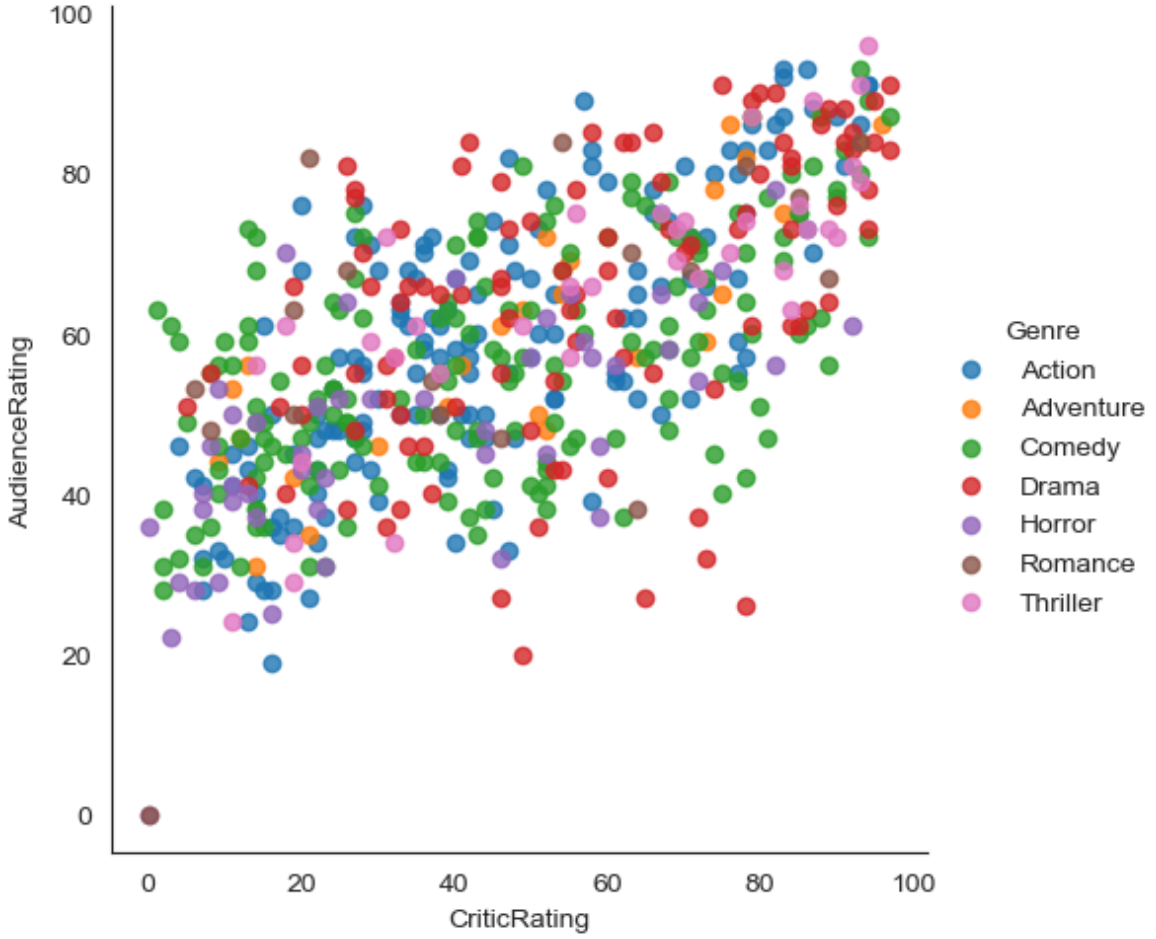
```
for gen in movies.Genre.cat.categories:
    print(gen)
```

```
Action
Adventure
Comedy
Drama
Horror
Romance
Thriller
```

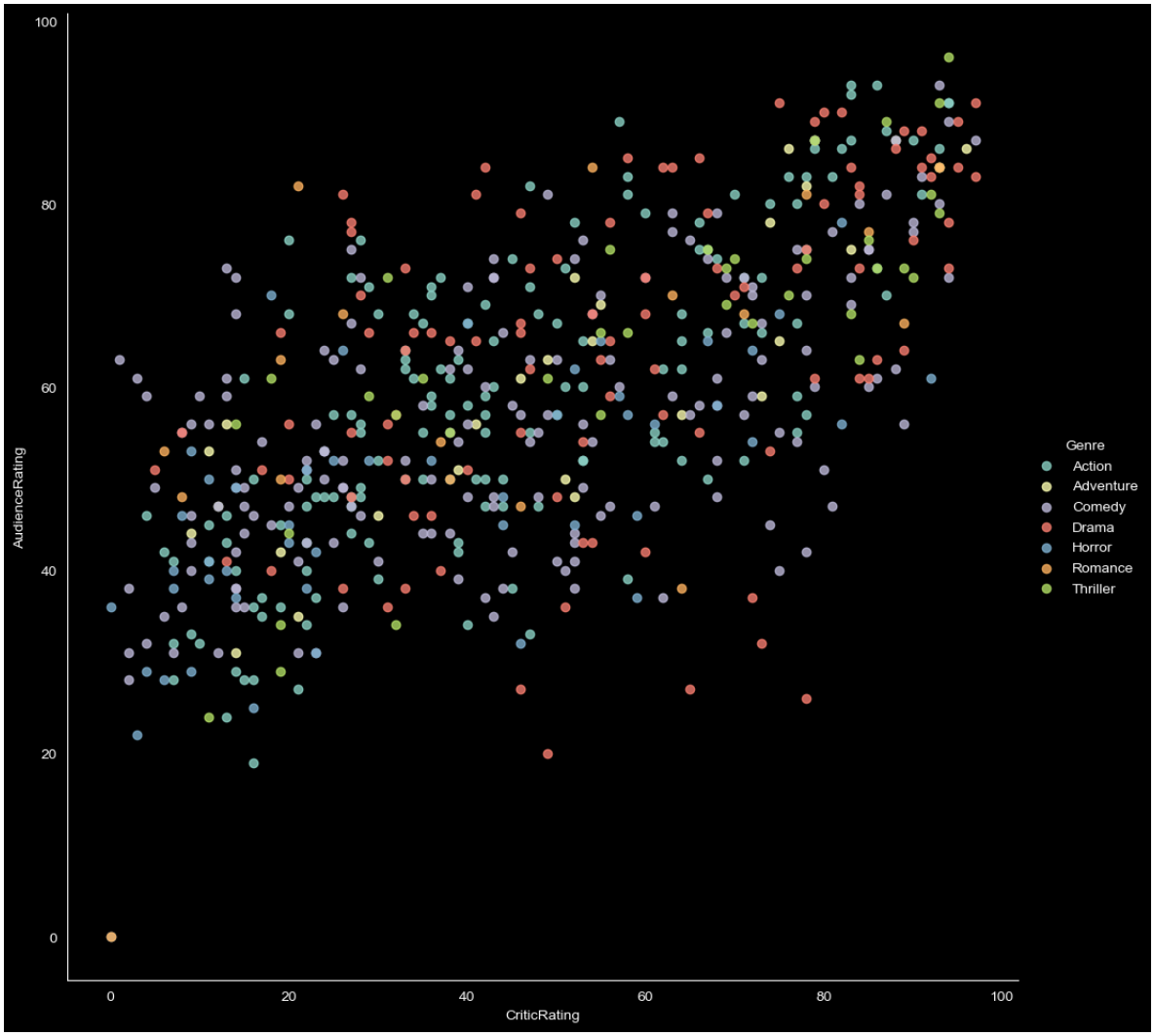
In [86]:
```
vis1 = sns.lmplot(data = movies, x = 'CriticRating', y = 'AudienceRating',\
                  fit_reg = False)
plt.show(vis1)
```



In [88]:
```
vis1 = sns.lmplot(data = movies, x = 'CriticRating', y = 'AudienceRating',\
                  fit_reg = False, hue = 'Genre')
plt.show(vis1)
```



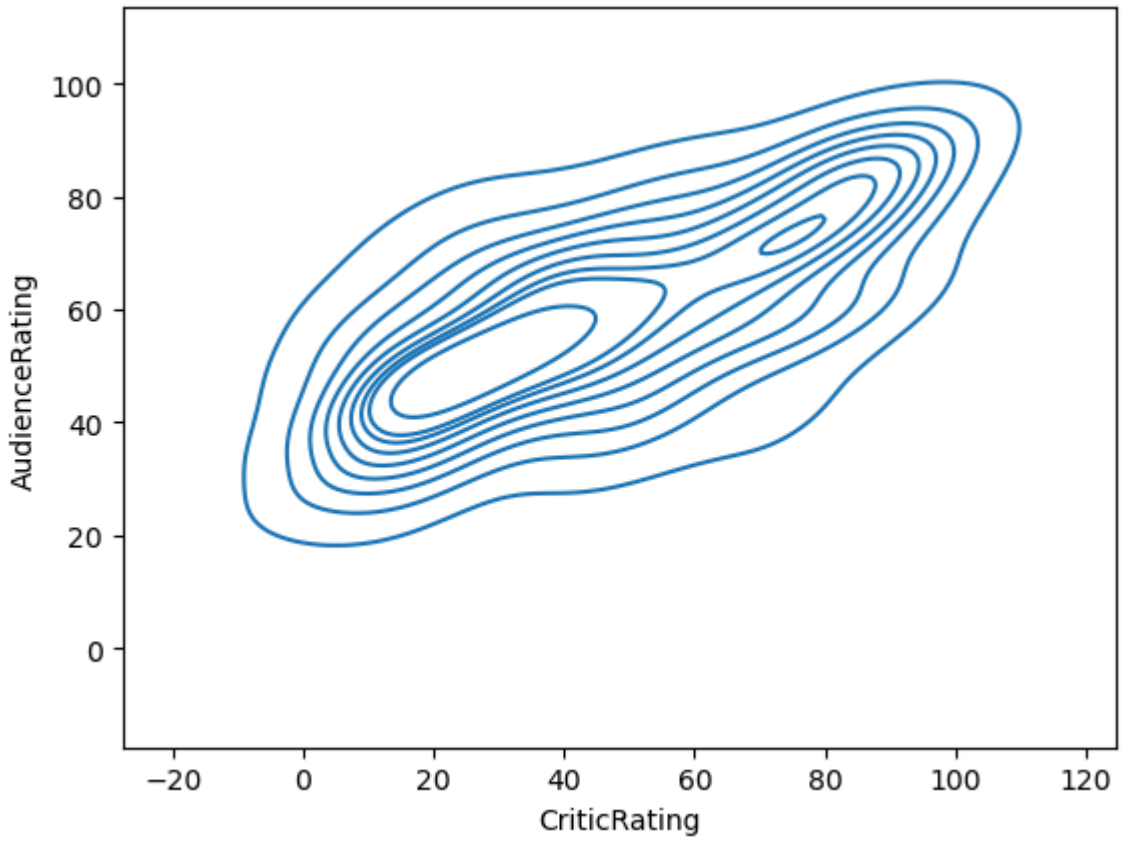
In [92]:
```
plt.style.use('dark_background')
vis2 = sns.lmplot(data=movies, x='CriticRating', y='AudienceRating', fit_reg=Fal
plt.show(vis2)
```



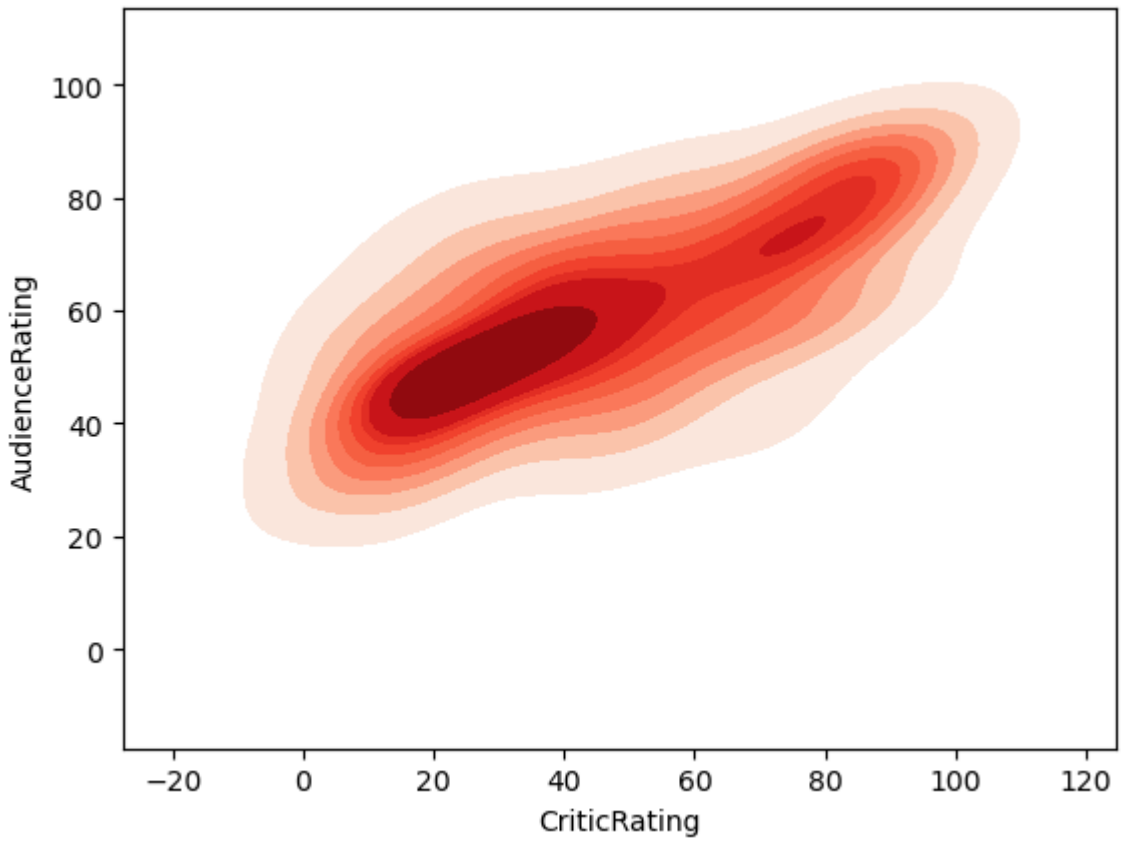
In [94]:
```
# Kernel Density Estimate plot (KDE PLOT)
# how can i visualize audiwece rating & critic Rating using scatterplot
```

In [96]:
```
plt.style.use('default')
```

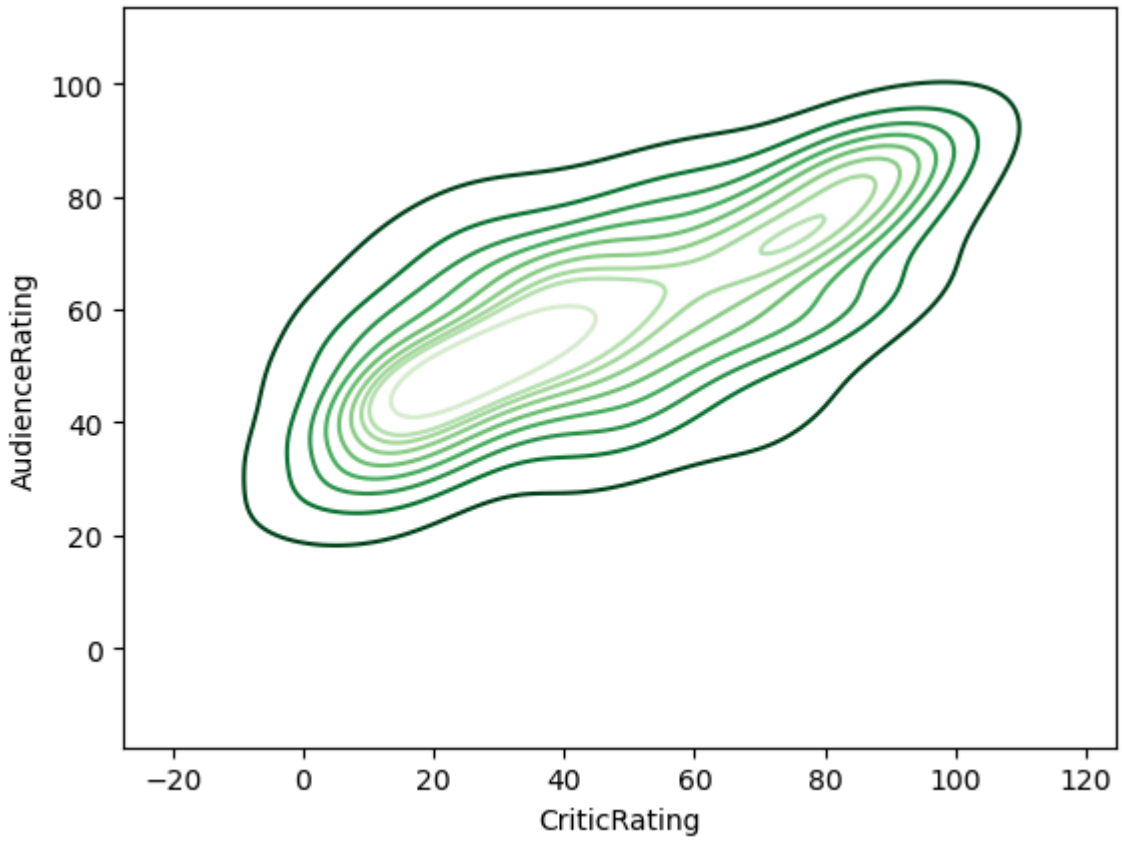
In [100…
```
k1 = sns.kdeplot(x = 'CriticRating', y = 'AudienceRating', data = movies)
plt.show(k1)
```



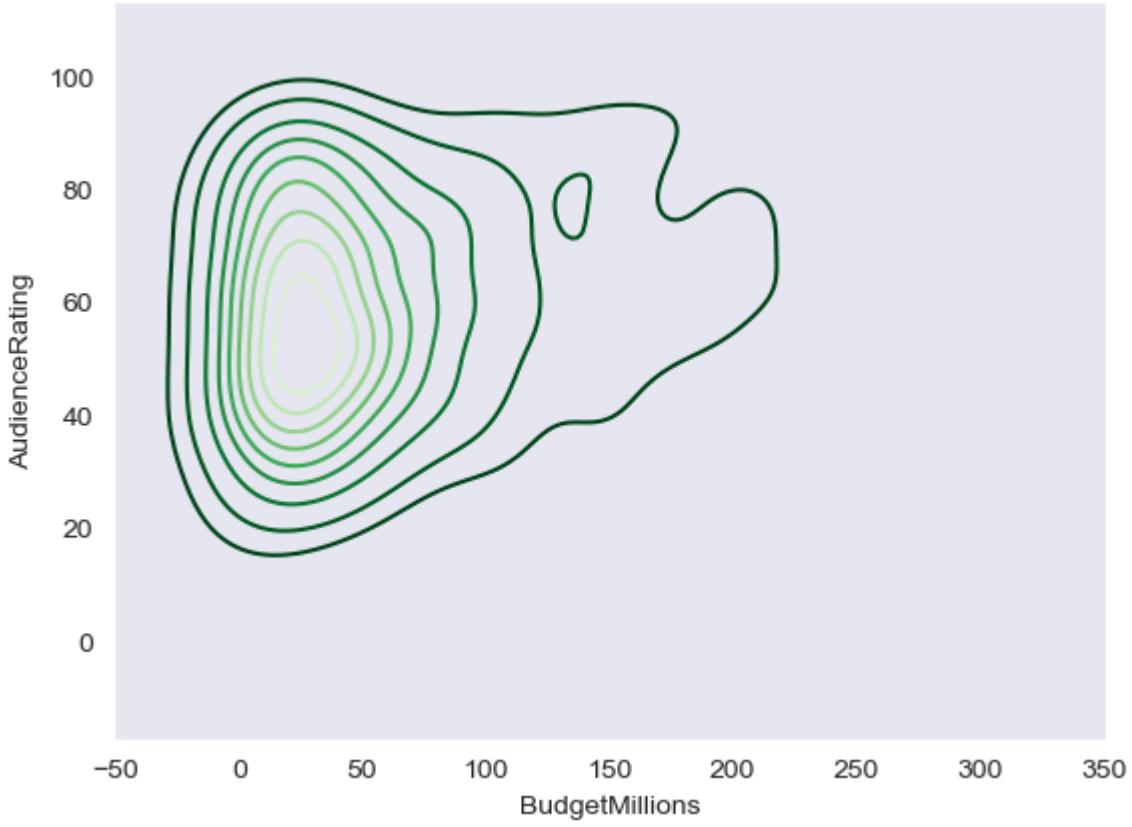
In [104…
```
k2 = sns.kdeplot(data = movies, x='CriticRating', y='AudienceRating', shade=True
plt.show(k2)
```



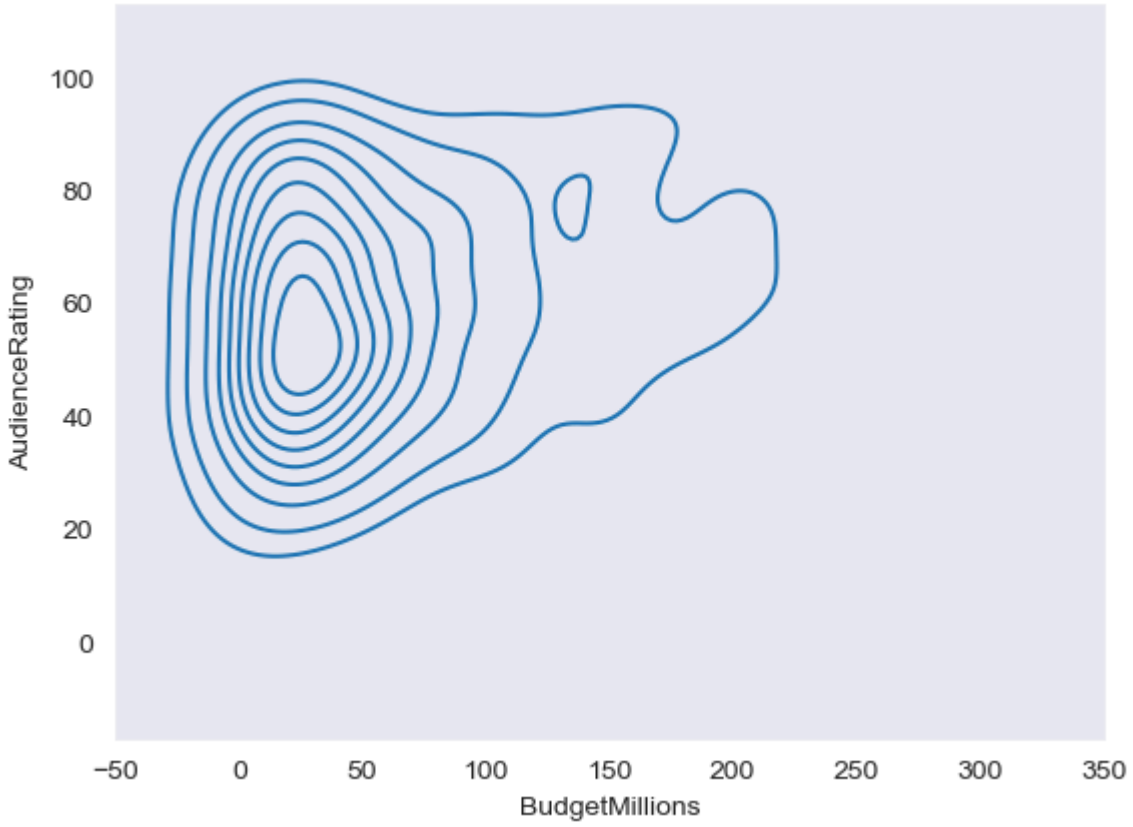
In [106…
```
k3 = sns.kdeplot(data = movies, x='CriticRating', y='AudienceRating', shade_lowe
plt.show(k3)
```



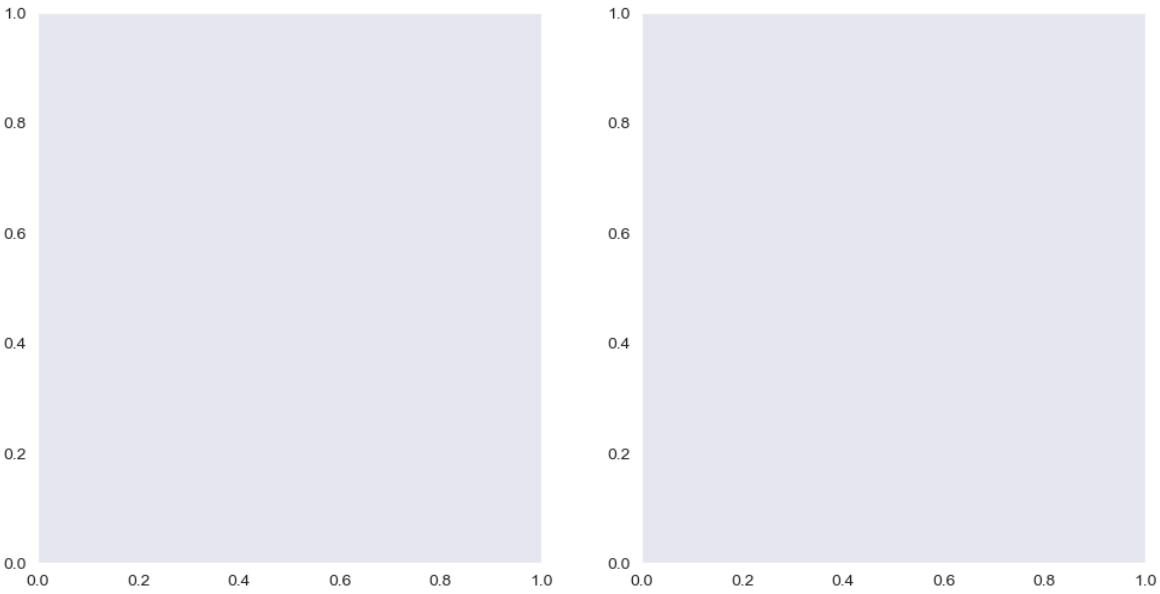
In [110…
```
sns.set_style('dark')
k4 = sns.kdeplot(data = movies, x='BudgetMillions', y='AudienceRating', shade_lo
plt.show(k4)
```



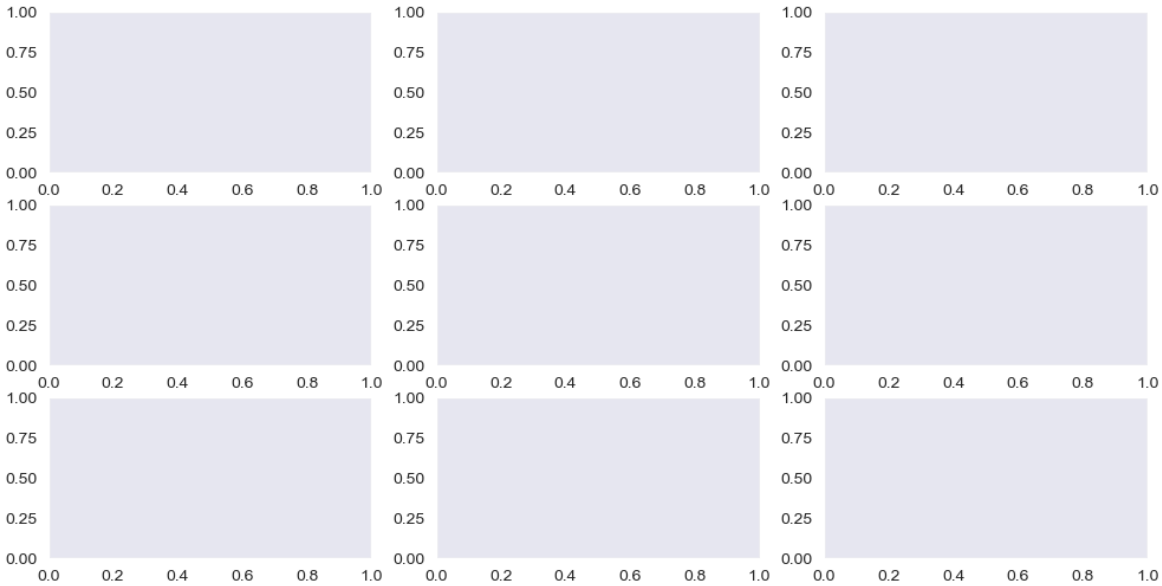
In [112…
```
sns.set_style('dark')
k5 = sns.kdeplot(data = movies, x='BudgetMillions', y='AudienceRating')
plt.show(k5)
```



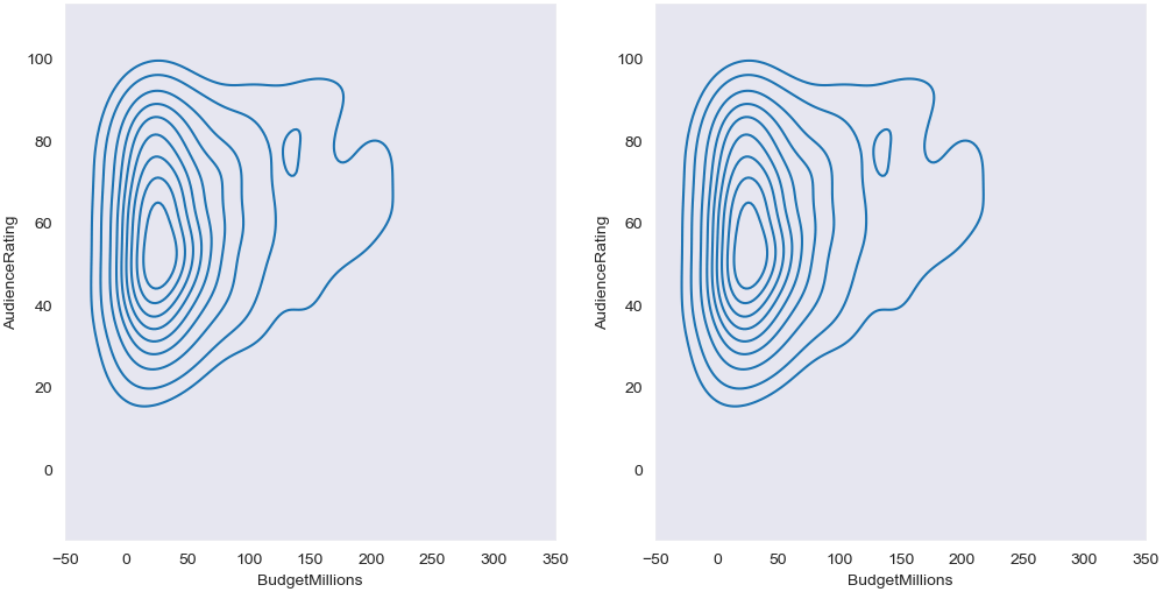
In [124…
```
# subplots

f, ax = plt.subplots(1,2, figsize =(12,6))
plt.show()
```



In [128…
```
f, ax = plt.subplots(3,3, figsize =(12,6))
plt.show()
```



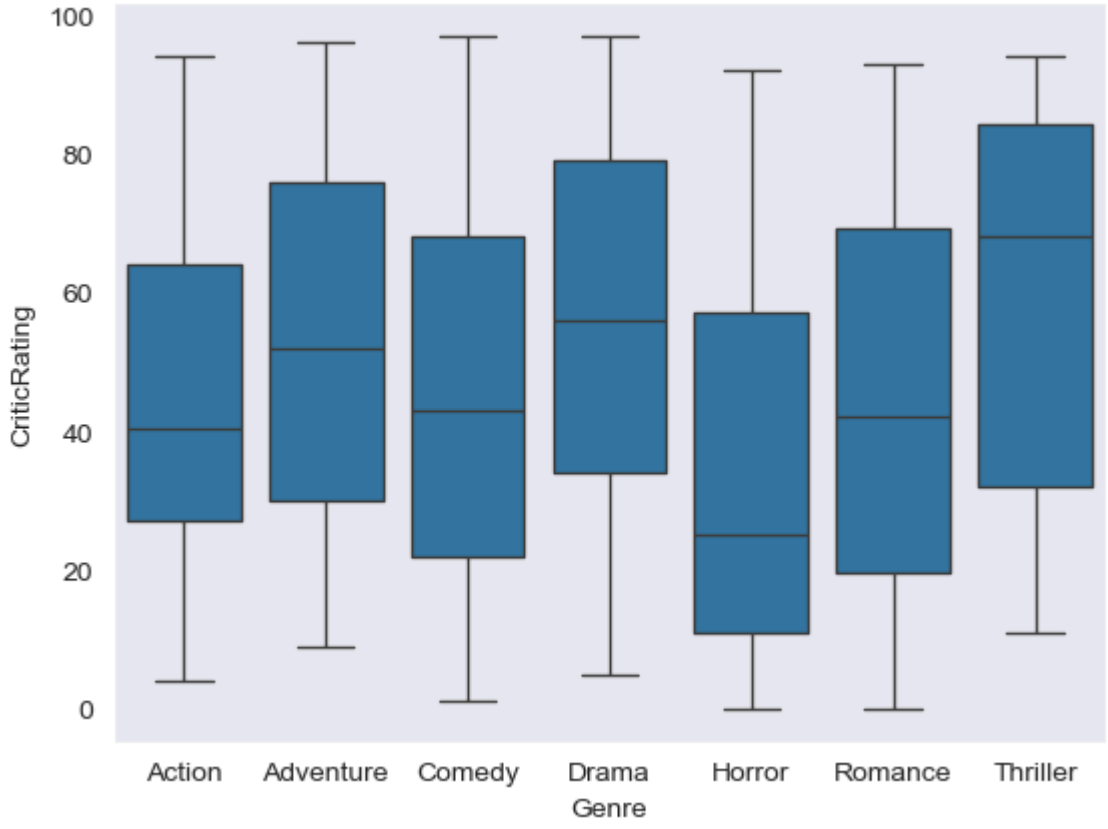
In [130…
```
f, axes = plt.subplots(1,2, figsize=(12,6))

k6 = sns.kdeplot(data = movies, x='BudgetMillions', y='AudienceRating', ax=axes[
k7 = sns.kdeplot(data = movies, x='BudgetMillions', y='AudienceRating', ax=axes[
plt.show()
```



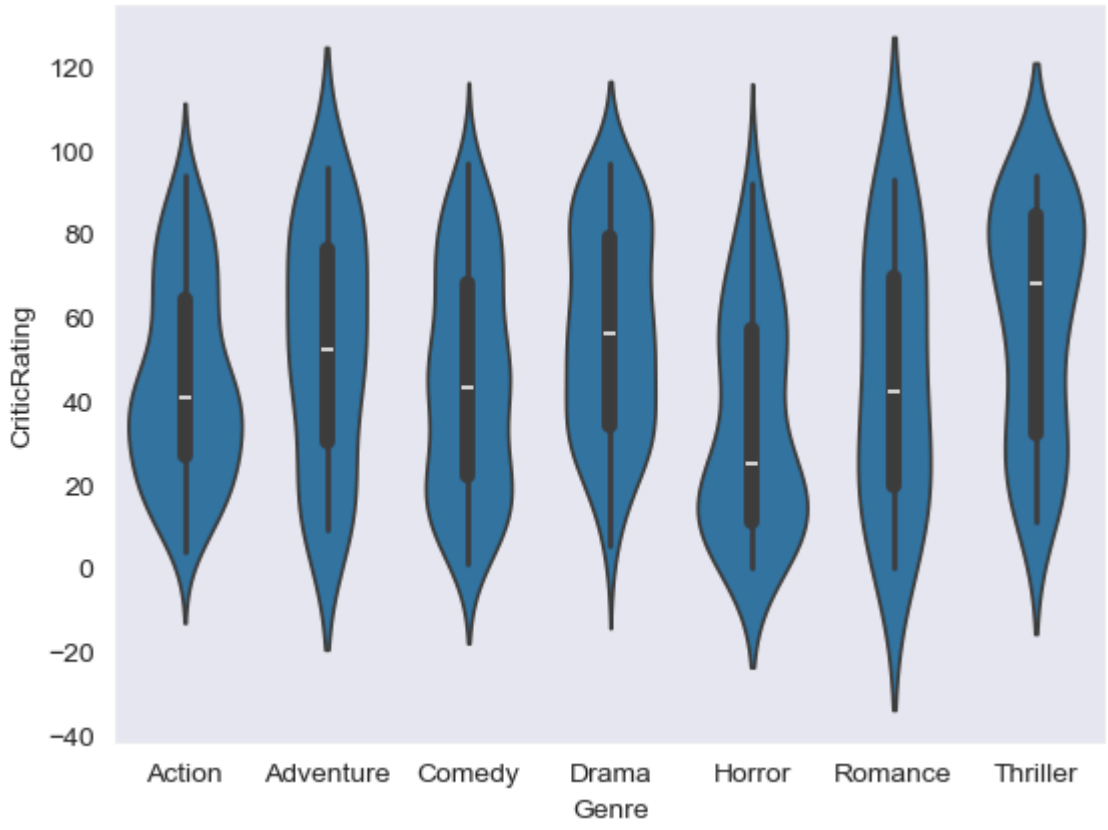
In [132…
```
axes
```

Out[132…
```
array([<Axes: xlabel='BudgetMillions', ylabel='AudienceRating'>,
       <Axes: xlabel='BudgetMillions', ylabel='AudienceRating'>],
      dtype=object)
```

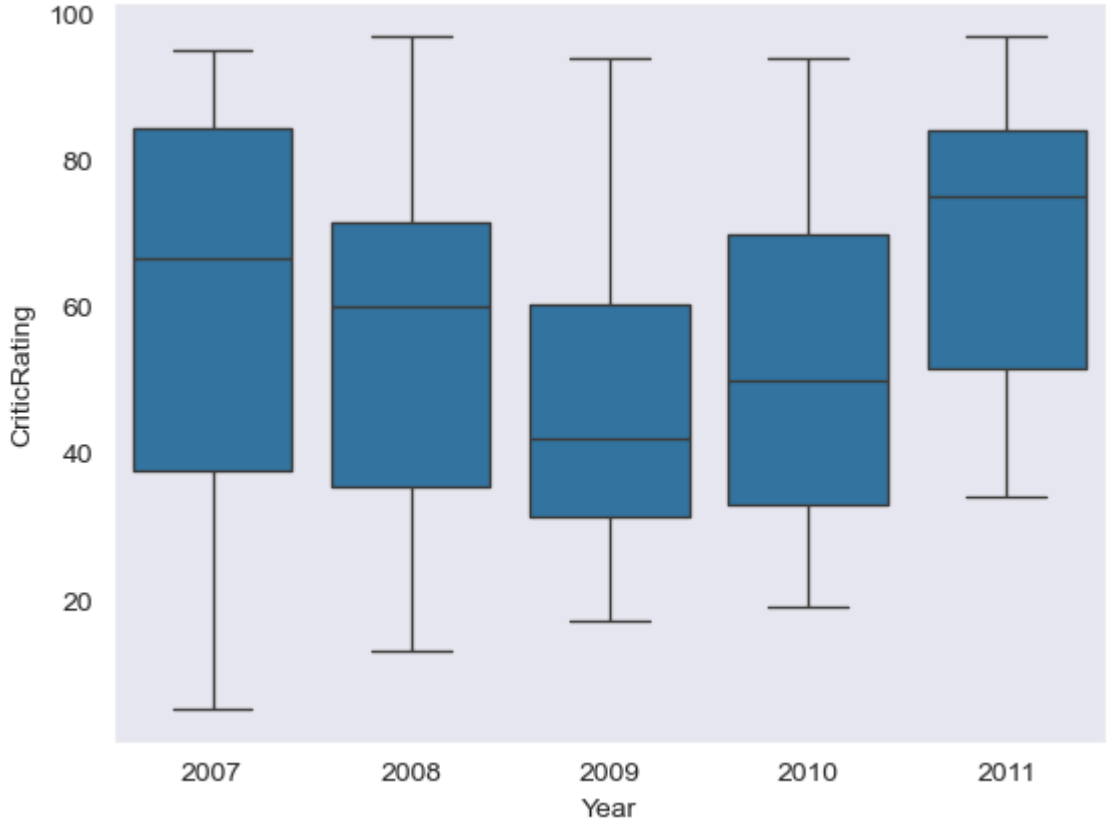
In [154…
```
# Box plots--

s = sns.boxplot(data = movies, x='Genre', y='CriticRating')
plt.show(s)
```



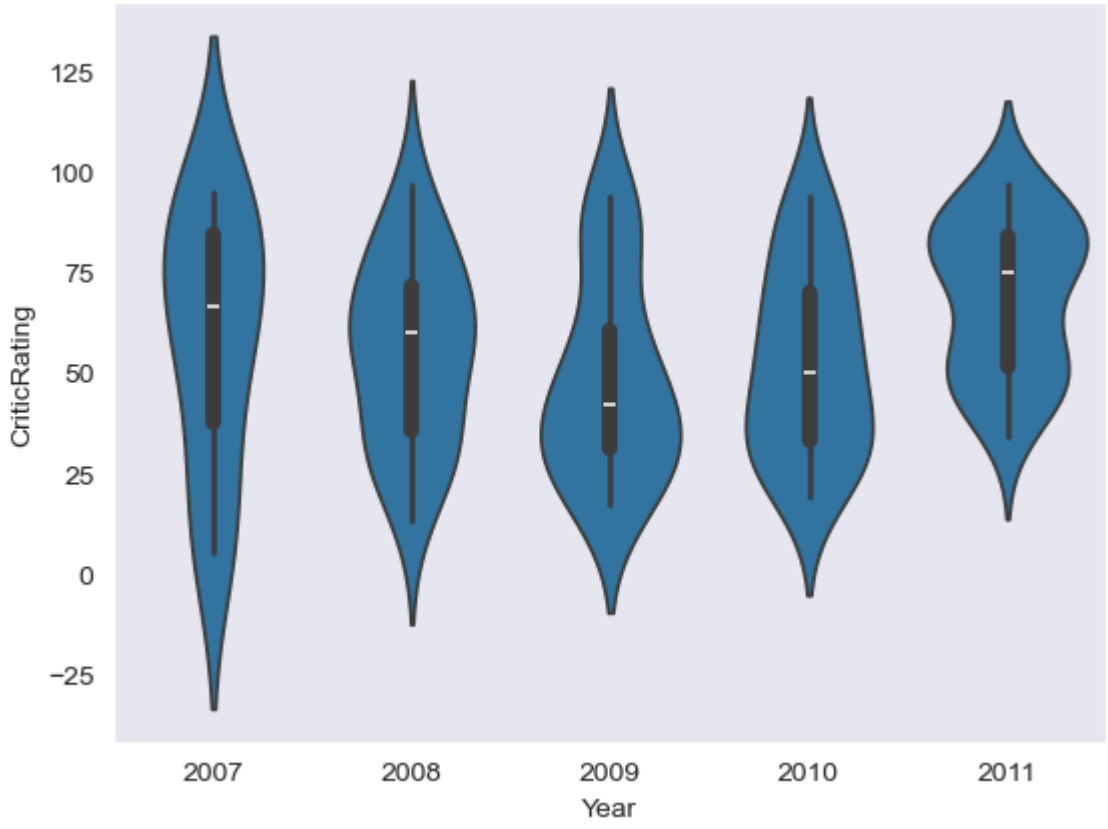
In [156…
```
#violin plot
s = sns.violinplot(data=movies, x='Genre', y = 'CriticRating')
plt.show(s)
```



In [162…
```
s1 = sns.boxplot(data=movies[movies.Genre == 'Drama'], x='Year', y = 'CriticRati
plt.show(s1)
```



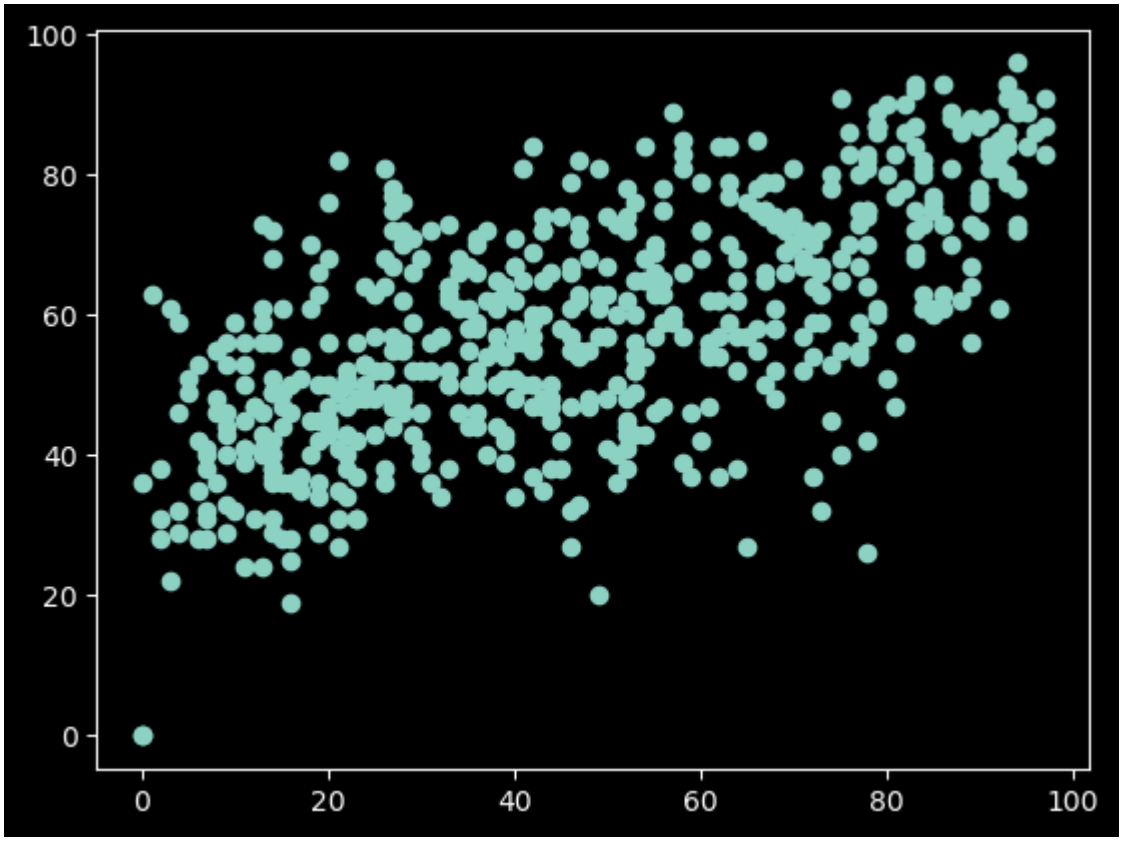
In [164…
```
z = sns.violinplot(data=movies[movies.Genre == 'Drama'], x='Year', y = 'CriticRa
plt.show(z)
```



In [166…
```
# Createing a Facet grid
plt.style.use('default')
```

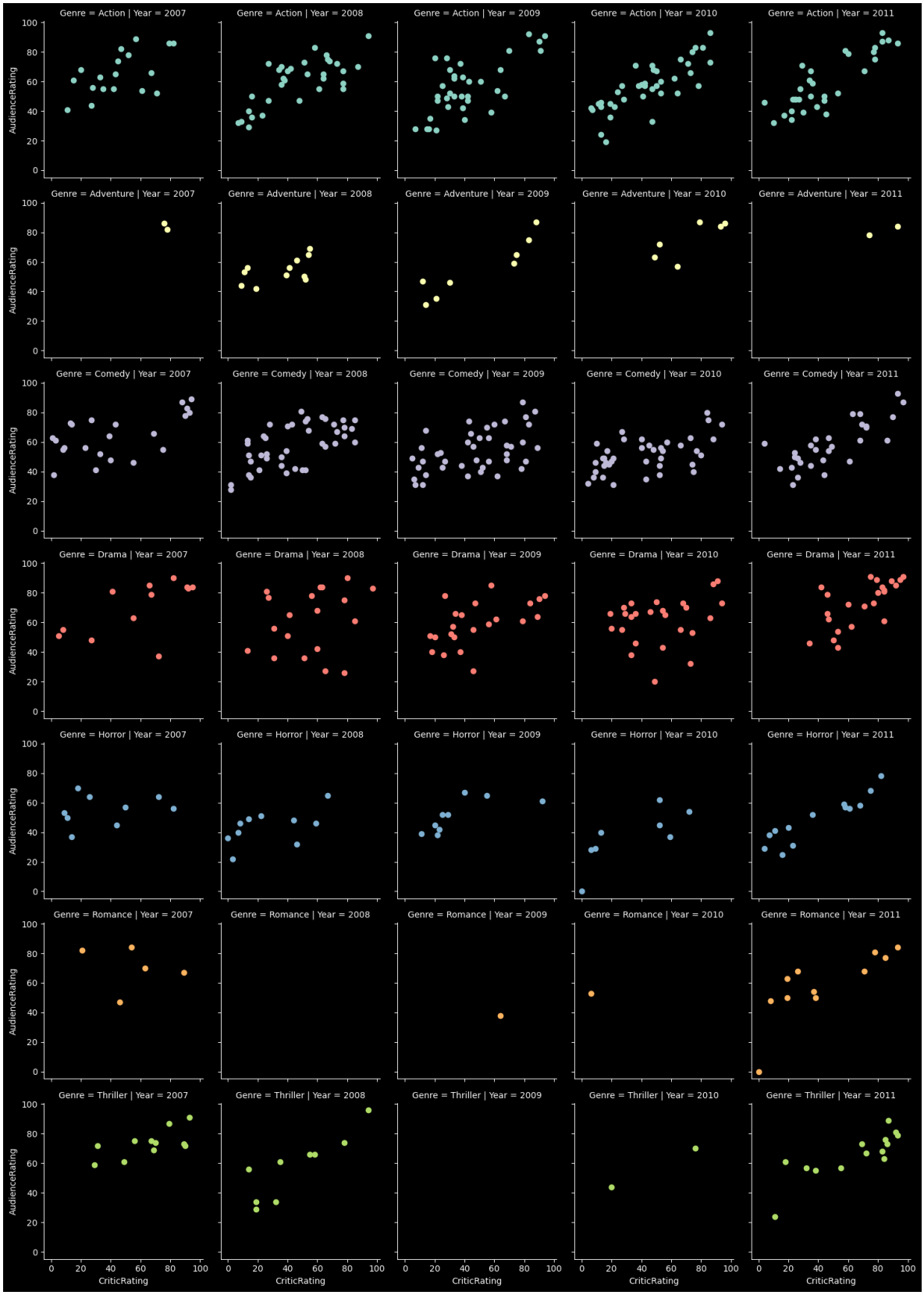
In [168…
```
g = sns.FacetGrid(movies, row ='Genre', col = 'Year', hue = 'Genre') #kind of su
plt.show(g)
```

In [178…
```
plt.style.use('dark_background')
plt.scatter(movies.CriticRating,movies.AudienceRating,)
plt.show()
```



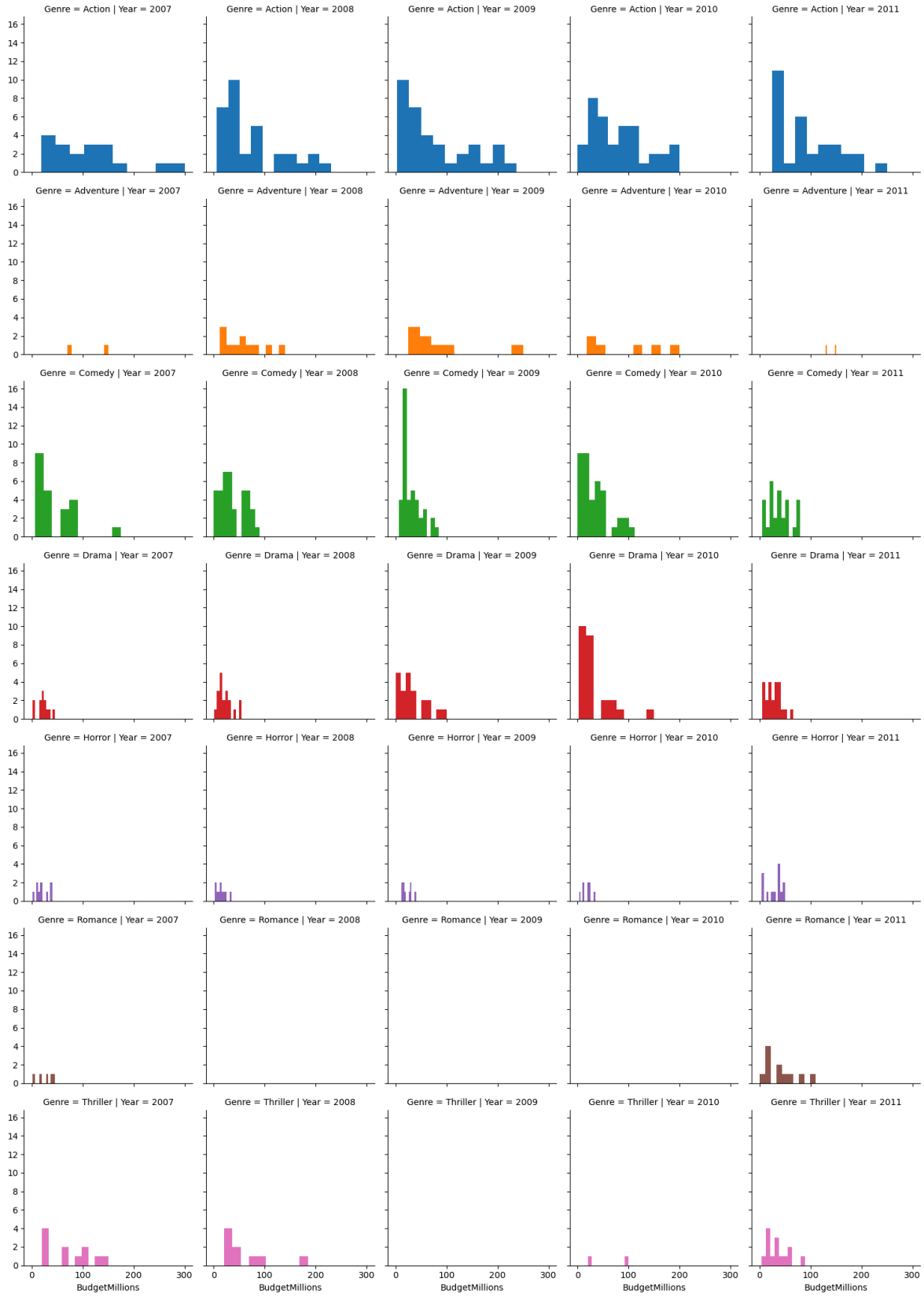
In [184…

```
plt.style.use('dark_background')
g =sns.FacetGrid (movies, row = 'Genre', col = 'Year', hue = 'Genre')
g = g.map(plt.scatter, 'CriticRating', 'AudienceRating' )
plt.show()
# scatterplots are mapped in facetgrid
```



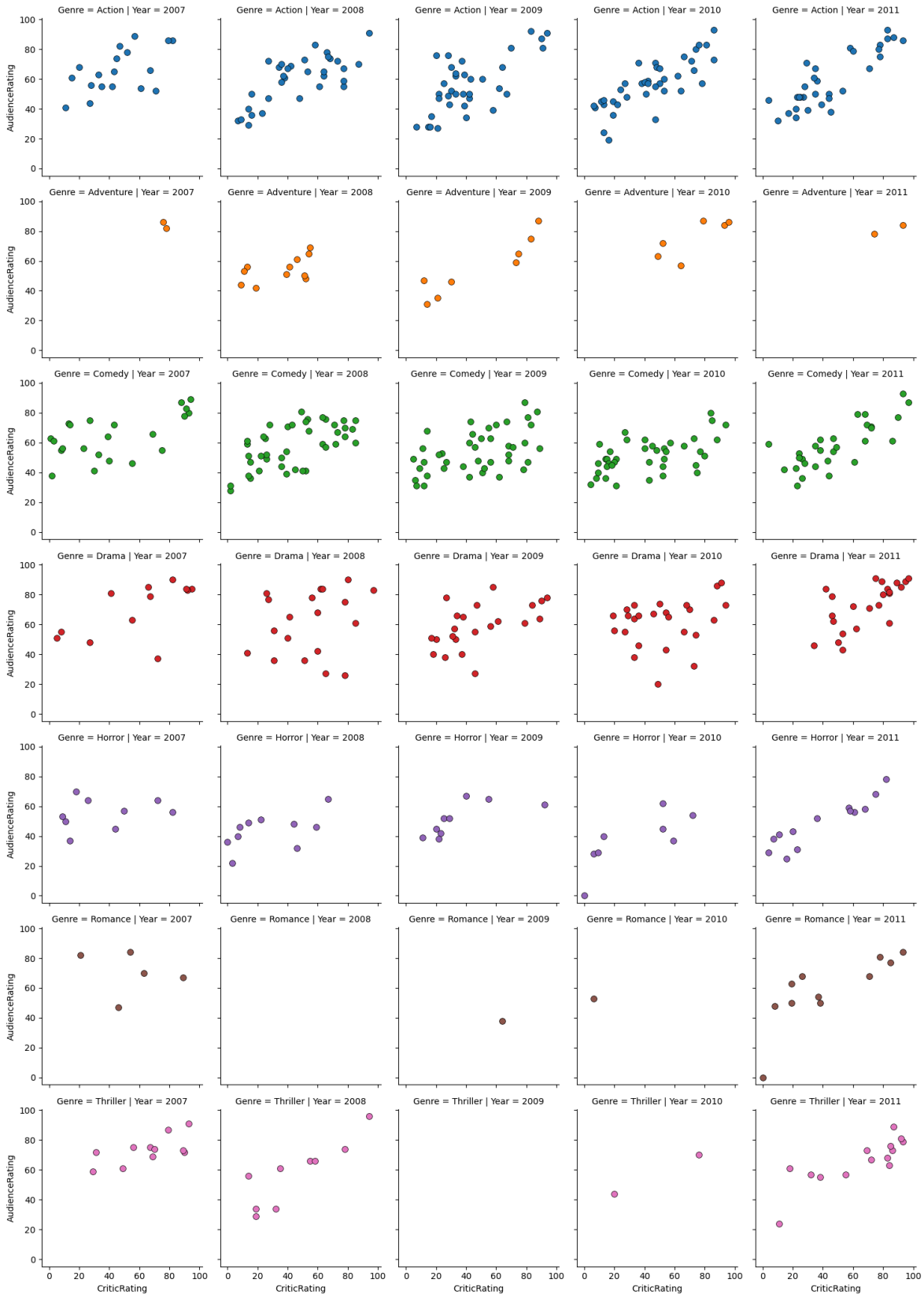
In [190…
```
plt.style.use('default')
```

In [192…
```
# you can populated any type of chat.

g = sns.FacetGrid (movies, row = 'Genre', col = 'Year', hue = 'Genre')
g = g.map(plt.hist, 'BudgetMillions') #scatterplots are mapped in facetgrid
plt.show()
```



In [194…

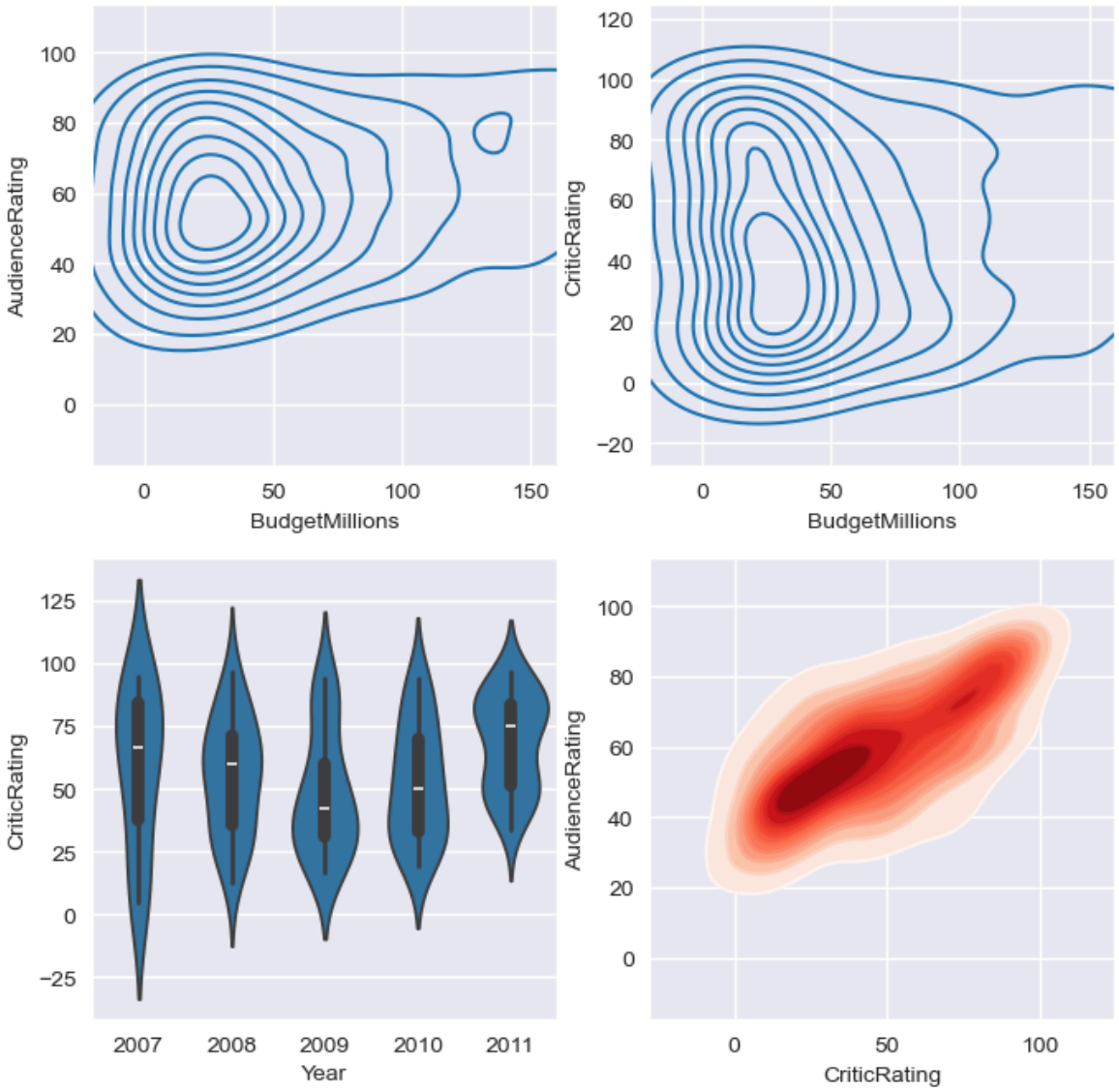
```
g = sns.FacetGrid(movies, row = 'Genre', col = 'Year', hue = 'Genre')
kws = dict(s=50, linewidth=0.5, edgecolor='black')
g = g.map(plt.scatter, 'CriticRating', 'AudienceRating', **kws) #scatterplots ar
plt.show()
```



In [196…

```
# python is not vectorize programming language
# Building dashboards (dashboards - combination of chats)

sns.set_style('darkgrid')
f, axes = plt.subplots (2,2, figsize = (8,8))


k1 = sns.kdeplot(x = movies.BudgetMillions,y = movies.AudienceRating,ax=axes[0,0
k2 = sns.kdeplot(x = movies.BudgetMillions,y = movies.CriticRating,ax = axes[0,1
```

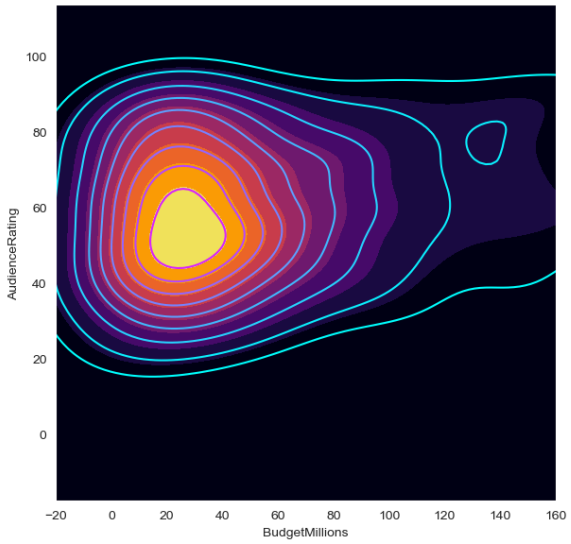
```
k1.set(xlim=(-20,160))
k2.set(xlim=(-20,160))

z = sns.violinplot(data=movies[movies.Genre=='Drama'], x='Year', y = 'CriticRati

k4 = sns.kdeplot(x = movies.CriticRating,y=movies.AudienceRating,shade = True,sh

k4b = sns.kdeplot(x = movies.CriticRating,y = movies.AudienceRating,cmap='Reds',

plt.show()
```

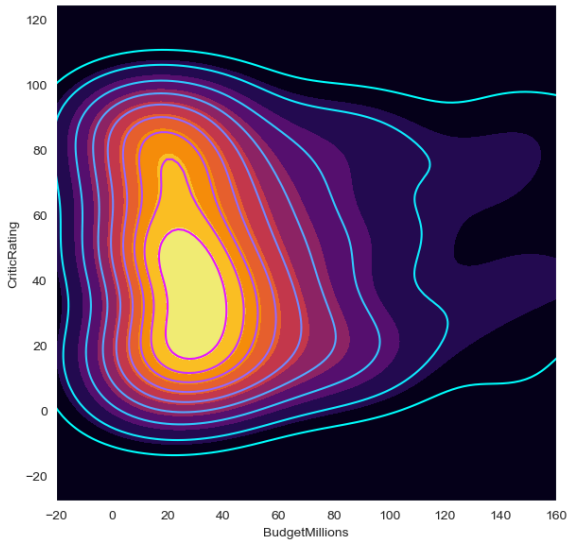


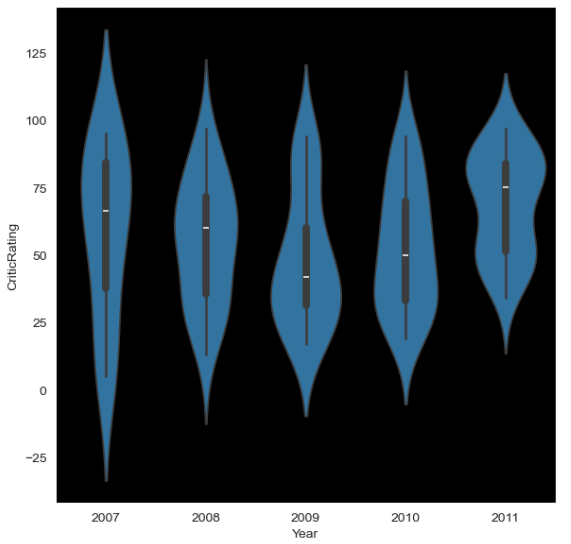
In [198…

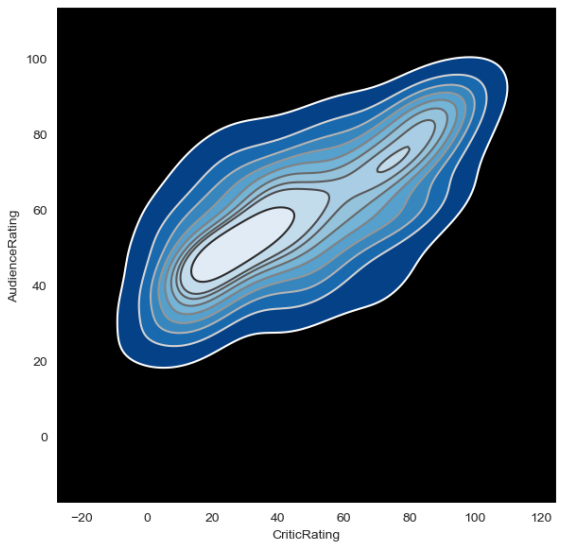
```
# How can you style your dashboard  using different color map

# python is not vectorize programming language
# Building dashboards (dashboard - combination of chats)

sns.set_style('dark',{'axes.facecolor':'black'})
f, axes = plt.subplots (2,2, figsize = (15,15))

#plot [0,0]
k1 = sns.kdeplot(x = movies.BudgetMillions,y = movies.AudienceRating, \
                 shade = True, shade_lowest=True,cmap = 'inferno', \
                 ax = axes[0,0])
k1b = sns.kdeplot(x= movies.BudgetMillions,y =  movies.AudienceRating, \
                 cmap = 'cool',ax = axes[0,0])
```

```
#plot [0,1]
k2 = sns.kdeplot(x= movies.BudgetMillions,y=movies.CriticRating,\
                 shade=True, shade_lowest=True, cmap='inferno',\
                 ax = axes[0,1])
k2b = sns.kdeplot(x=movies.BudgetMillions,y=movies.CriticRating,\
                  cmap = 'cool', ax = axes[0,1])

#plot[1,0]
z = sns.violinplot(data=movies[movies.Genre=='Drama'], \
                   x='Year', y = 'CriticRating', ax=axes[1,0])

#plot[1,1]
k4 = sns.kdeplot(x = movies.CriticRating,y =movies.AudienceRating, \
                 shade = True,shade_lowest=False,cmap='Blues_r', \
                 ax=axes[1,1])

k4b = sns.kdeplot(x = movies.CriticRating,y = movies.AudienceRating, \
                  cmap='gist_gray_r',ax = axes[1,1])


k1.set(xlim=(-20,160))
k2.set(xlim=(-20,160))

plt.show()
```









Final discussion what we learn so far - 1> category datatype in python 2> jointplots 3> histogram 4> stacked histograms 5> Kde plot 6> subplot 7> violin plots 8> Factet grid 9> Building dashboards

In [ ]: