In [56]:
```
# MOVIE RATING ANALYTICS (ADVANCED VISULIZATION)

import pandas as pd
import os
```

In [57]:
```
os.getcwd()
```

Out[57]:
```
'c:\\Users\\admin\\VSCODE'
```

In [58]:
```
movies=pd.read_csv(r"C:\Users\admin\Desktop\Data Science 7pm\3rd Sep-MOVIE RATINGS
```

In [59]:
```
movies
```

Out[59]:

	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 [60]:
```
len(movies)
```

Out[60]:
```
559
```

In [61]:
```
movies.head()
```

Out[61]:

	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 [62]:
```
movies.tail()
```

Out[62]:

	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 [63]:
```
movies.columns
```

Out[63]:
```
Index(['Film', 'Genre', 'Rotten Tomatoes Ratings %', 'Audience Ratings %',
       'Budget (million $)', 'Year of release'],
      dtype='object')
```

In [65]:
```
movies.columns = ['Film', 'Genre', 'CriticRating', 'AudienceRating','BudgetMillions
```

In [46]:
```
movies.head()
```

Out[46]:

	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 [47]:
```
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   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: category(2), int64(4)
memory usage: 40.1 KB
```

In [48]:
```
movies.describe()
# if you look at the year the data type is int but when you look at the mean value
# we have to change to categroy type
# also from object datatype we will convert to category datatypes
#
```

Out[48]:

	Rotten Tomatoes Ratings %	Audience Ratings %	Budget (million $)	Year of release
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 [49]:
```
movies['Film']
#movies['Audience Ratings %']
```

Out[49]:
```
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 [50]:
```
movies.Film
```

Out[50]:
```
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 [51]:
```
movies.Film = movies.Film.astype('category')
```

In [52]:
```
movies.Film
```

Out[52]:
```
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 [53]:
```
movies.head()
```

Out[53]:

	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 [54]:
```
movies.info()

# now the same thing we will change genra to category & year to category
```

```
<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   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: category(2), int64(4)
memory usage: 40.1 KB
```

In [66]:
```
movies.Genre = movies.Genre.astype('category')
movies.Year = movies.Year.astype('category')
```

In [67]:
```
movies.Genre
```

Out[67]:
```
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', 'Roma
nce', 'Thriller']
```

In [68]:
```
movies.Year
```

Out[68]:
```
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 [69]:
```
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    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(2), int64(3), object(1)
memory usage: 19.2+ KB
```

In [70]:
```
movies.Genre.cat.categories
```

Out[70]:
```
Index(['Action', 'Adventure', 'Comedy', 'Drama', 'Horror', 'Romance',
       'Thriller'],
      dtype='object')
```

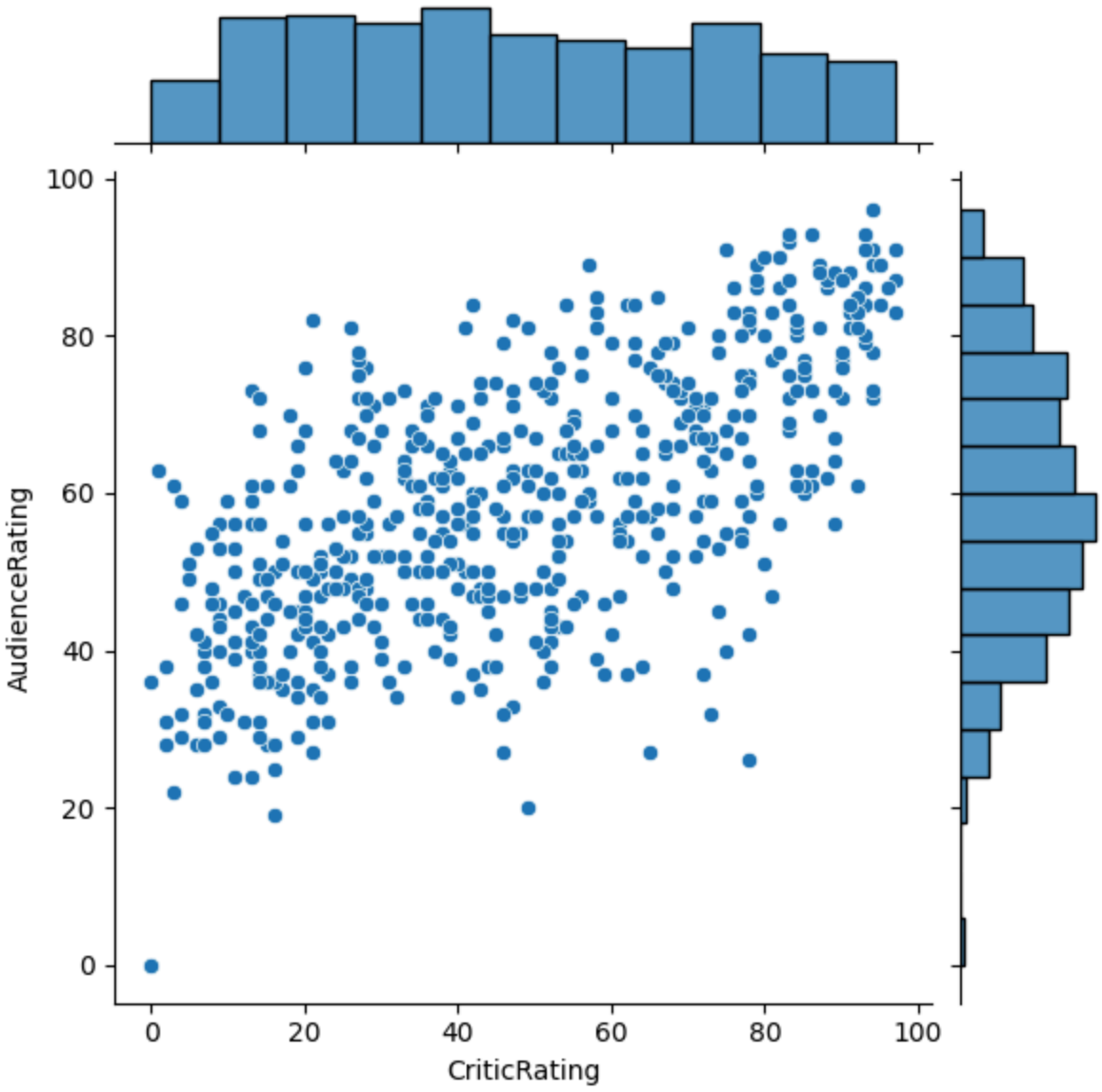
In [71]:
```
movies.describe()
#now when you see the describt you will get only integer value mean, standard devia
```

Out[71]:

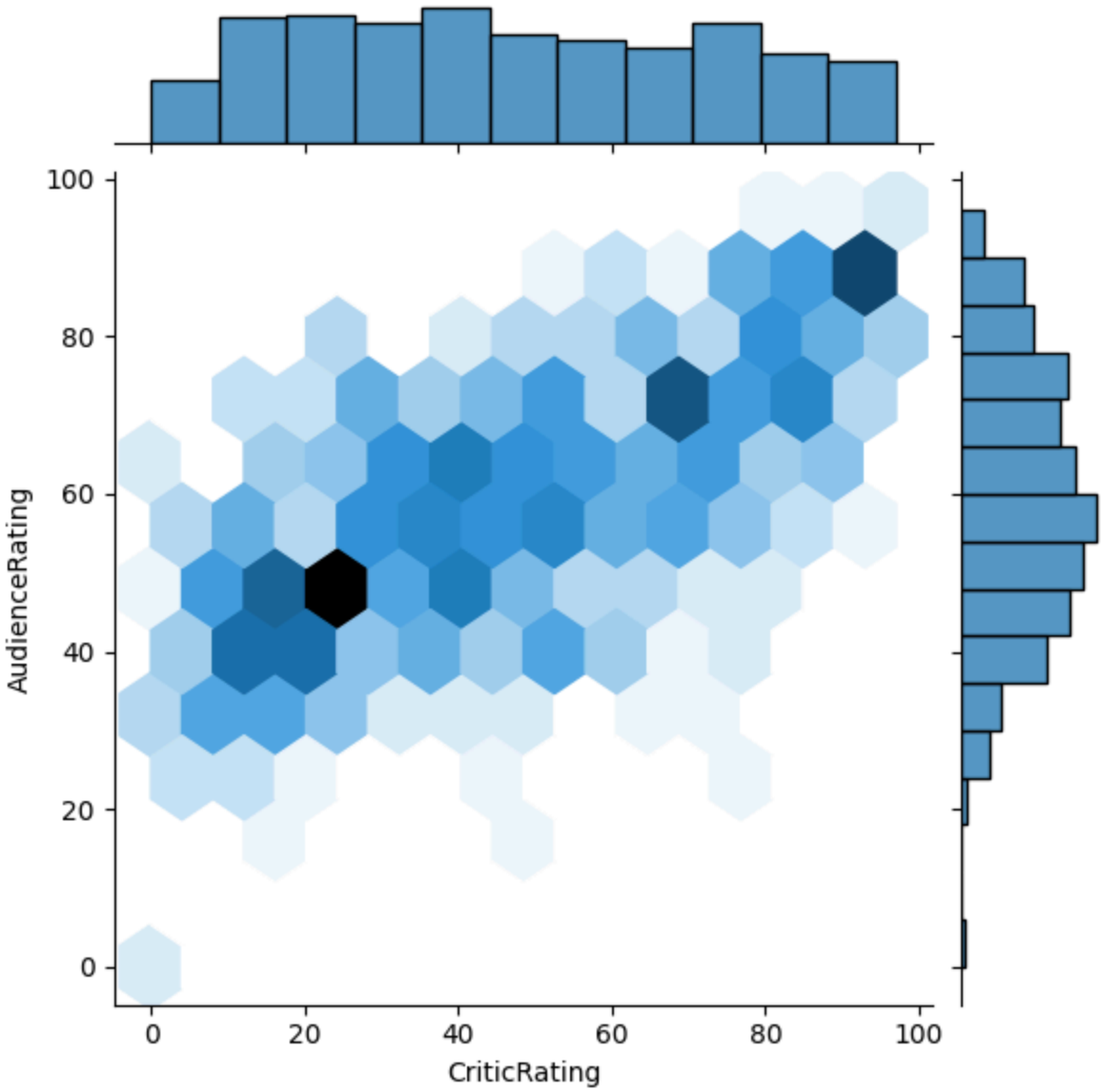
	CriticRating	AudienceRating	BudgetMillions
count	559.000000	559.000000	559.000000
mean	47.309481	58.744186	50.236136
std	26.413091	16.826887	48.731817
min	0.000000	0.000000	0.000000
25%	25.000000	47.000000	20.000000
50%	46.000000	58.000000	35.000000
75%	70.000000	72.000000	65.000000
max	97.000000	96.000000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

In [72]:
```
# How to working with joint plots

from matplotlib import pyplot as plt
import seaborn as sns
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

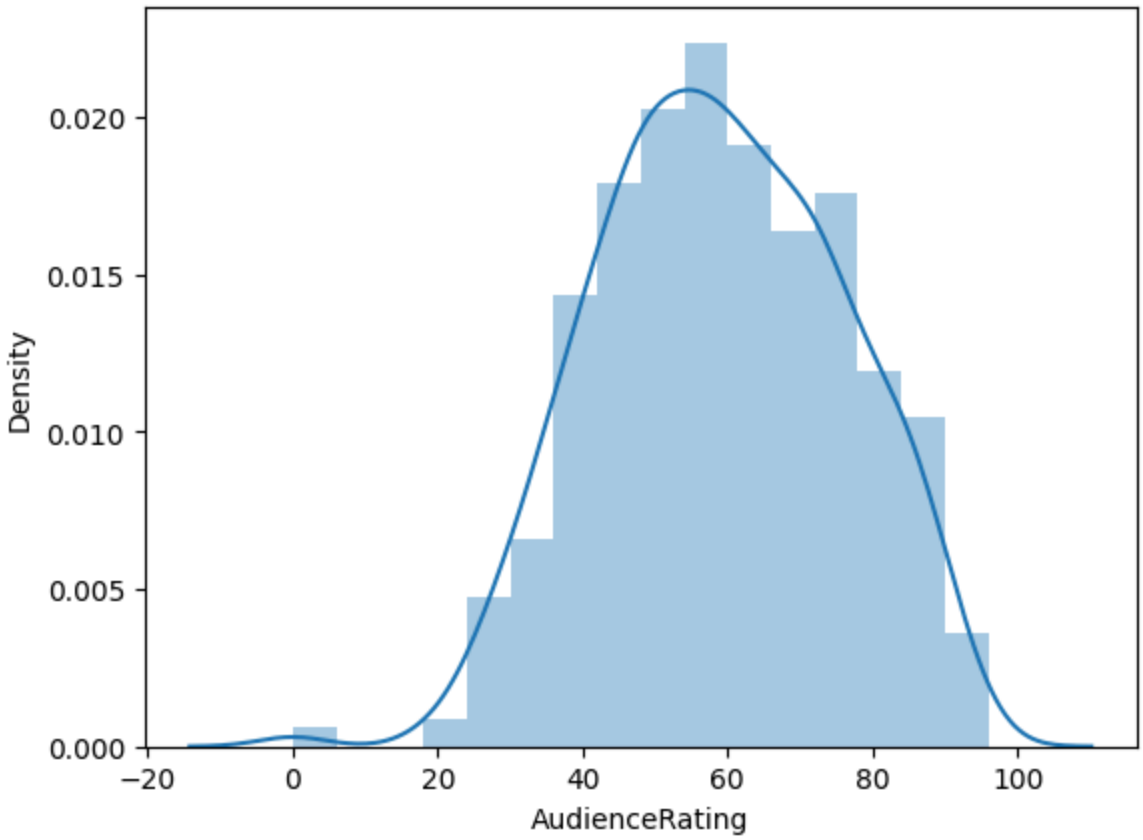
In [73]:
```
j = sns.jointplot( data = movies, x = 'CriticRating', y = 'AudienceRating')
# Audience rating is more dominant then critics rating
# Based on this we find out as most people are most liklihood to watch audience rat
# let me explain the excel - if you filter audience rating & critic rating. critic
```



In [74]:
```
j = sns.jointplot( data = movies, x = 'CriticRating', y = 'AudienceRating', kind='h

#j = sns.jointplot( data = movies, x = 'CriticRating', y = 'AudienceRating', kind='
```

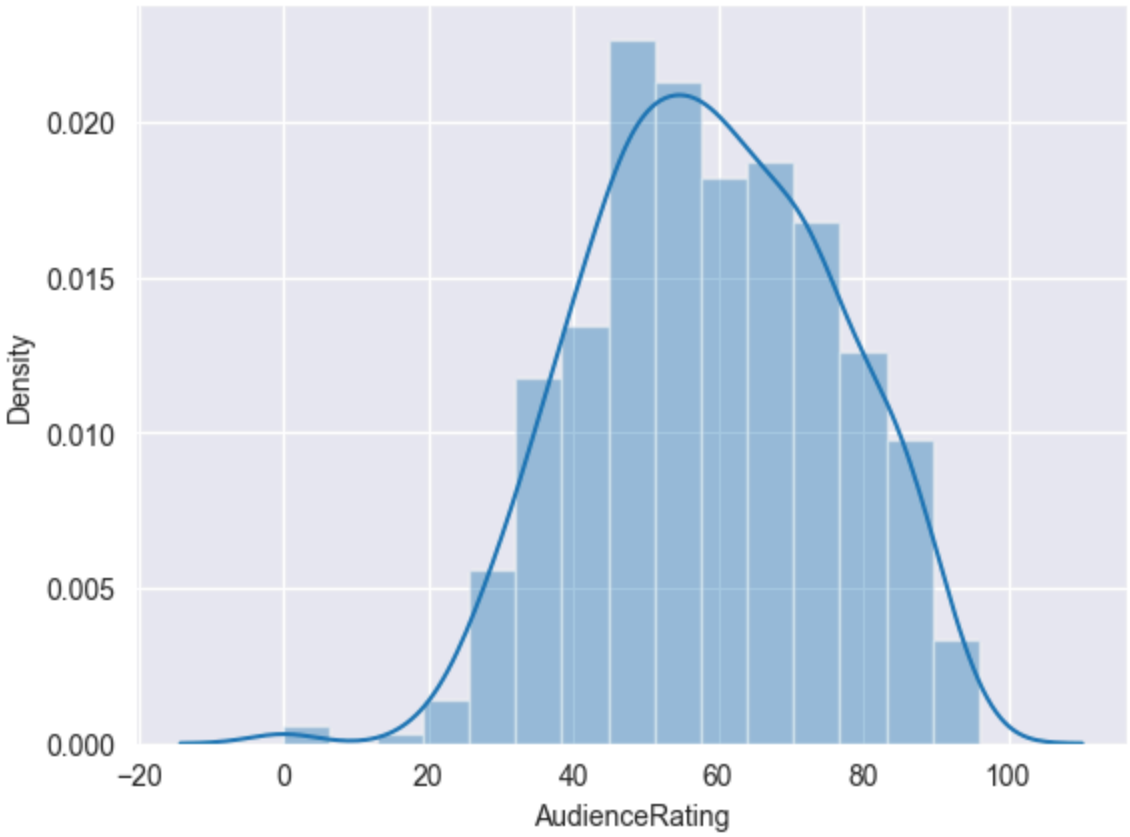


In [75]:
```
#Histograms

# <<< chat1

m1 = sns.distplot(movies.AudienceRating)

#y - axis generated by seaborn automatically that is the powefull of seaborn galler
```

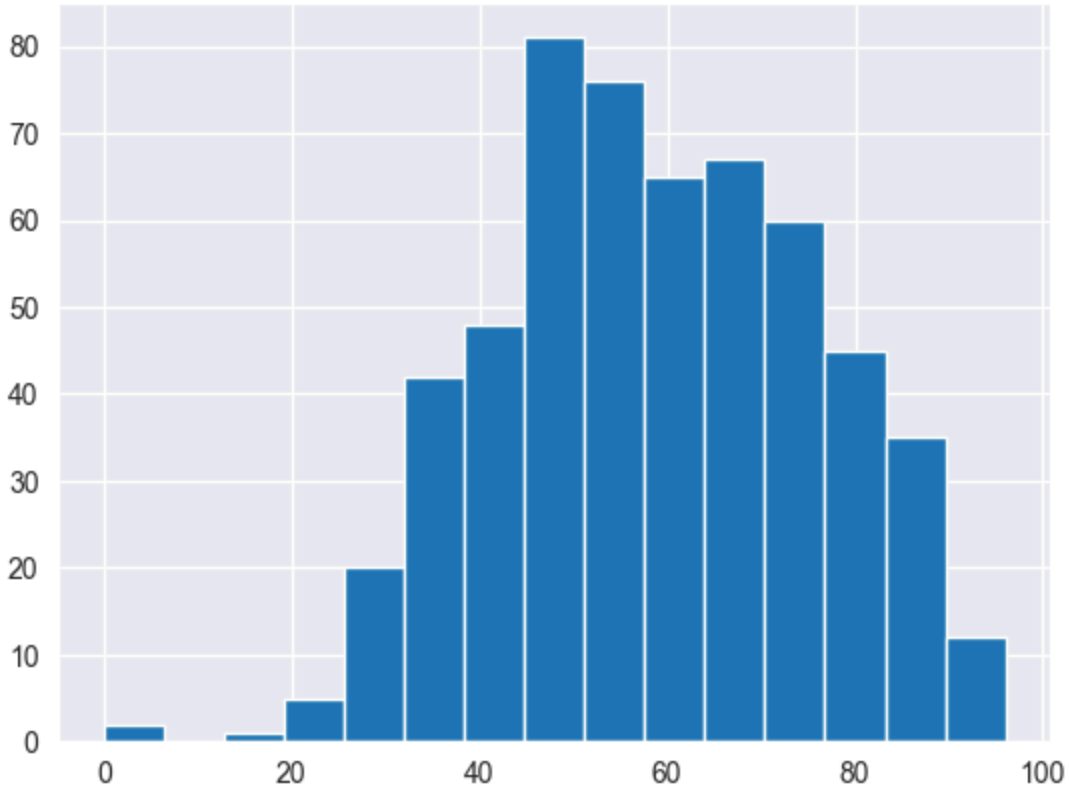


In [76]:
```
sns.set_style('darkgrid')
```

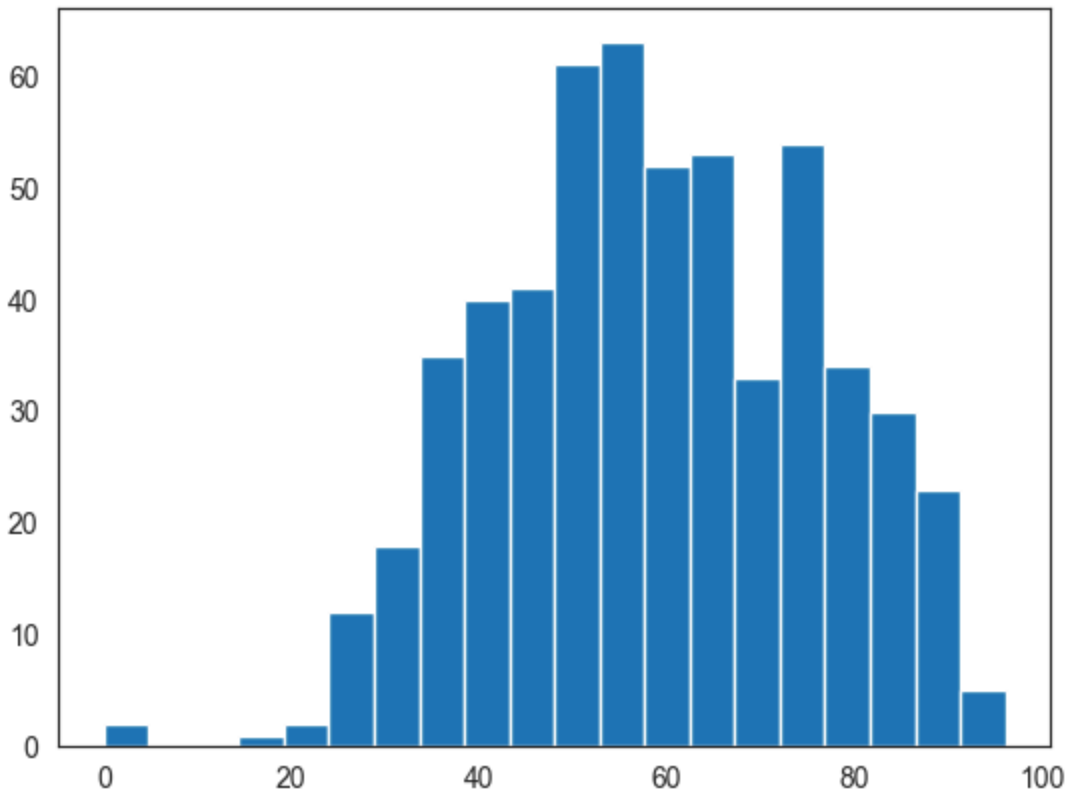
In [77]:
```
m2 = sns.distplot(movies.AudienceRating, bins = 15)
```



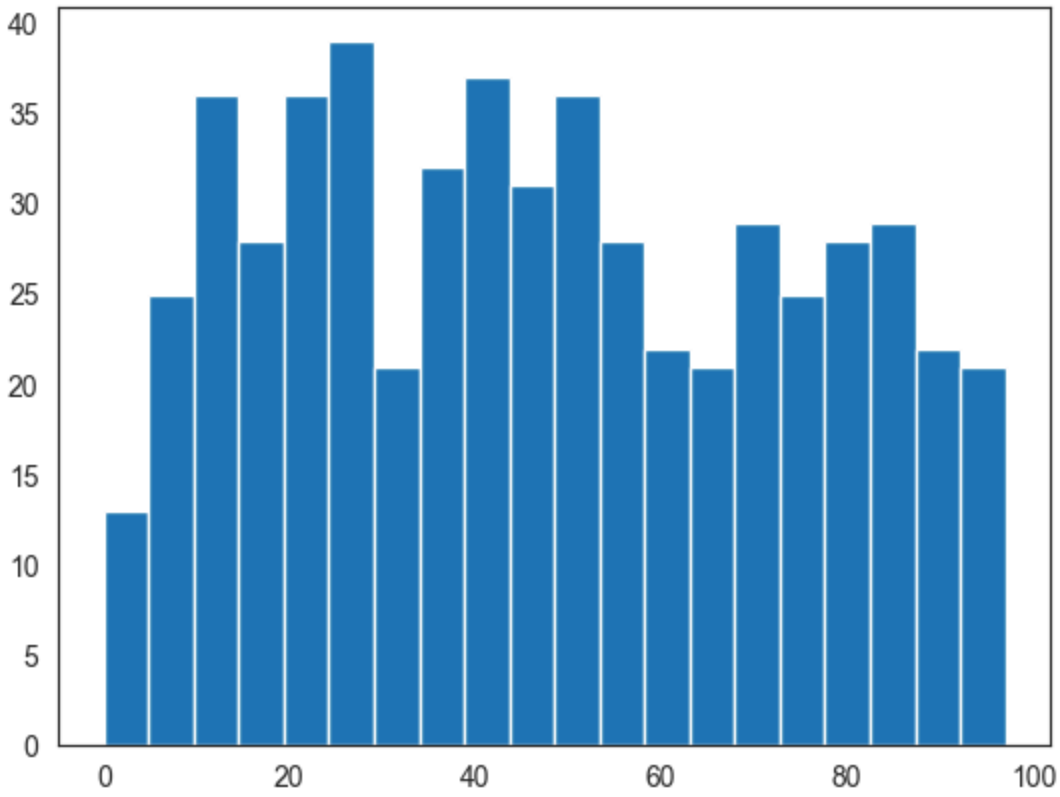
In [78]:
```
#sns.set_style('darkgrid')
n1 = plt.hist(movies.AudienceRating, bins=15)
```



In [79]:
```
sns.set_style('white') #normal distribution & called as bell curve
n1 = plt.hist(movies.AudienceRating, bins=20)
```

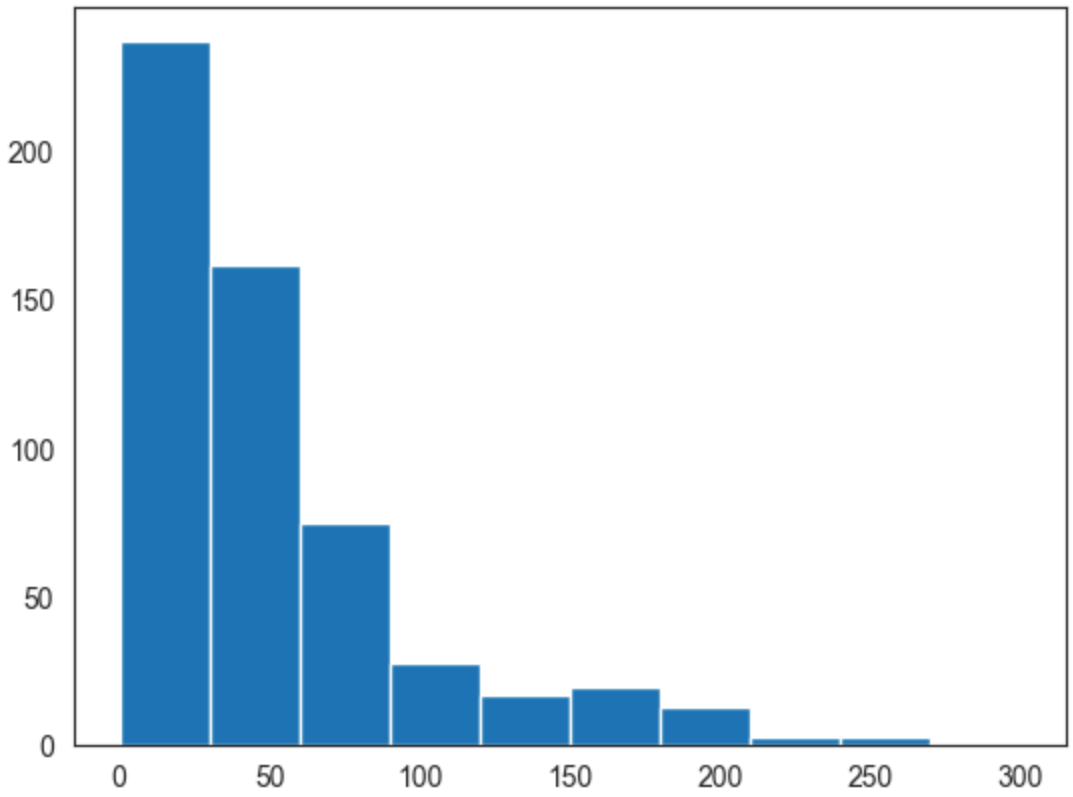


In [80]:
```
n1 = plt.hist(movies.CriticRating, bins=20) #uniform distribution
```

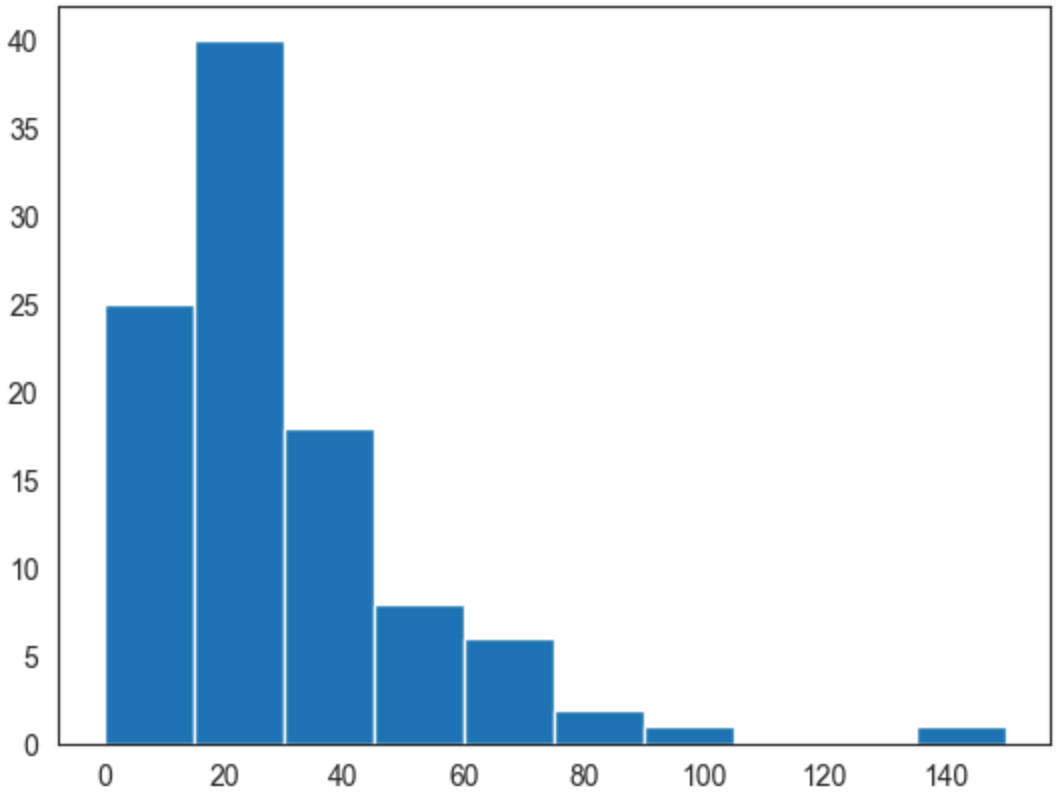


In [81]:
```
# <<< chat - 2

# Creating stacked histograms & this is bit tough to understand
```

In [82]:
```
#h1 = plt.hist(movies.BudgetMillions)

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



In [83]:
```
plt.hist(movies[movies.Genre == 'Drama'].BudgetMillions)
plt.show()
```



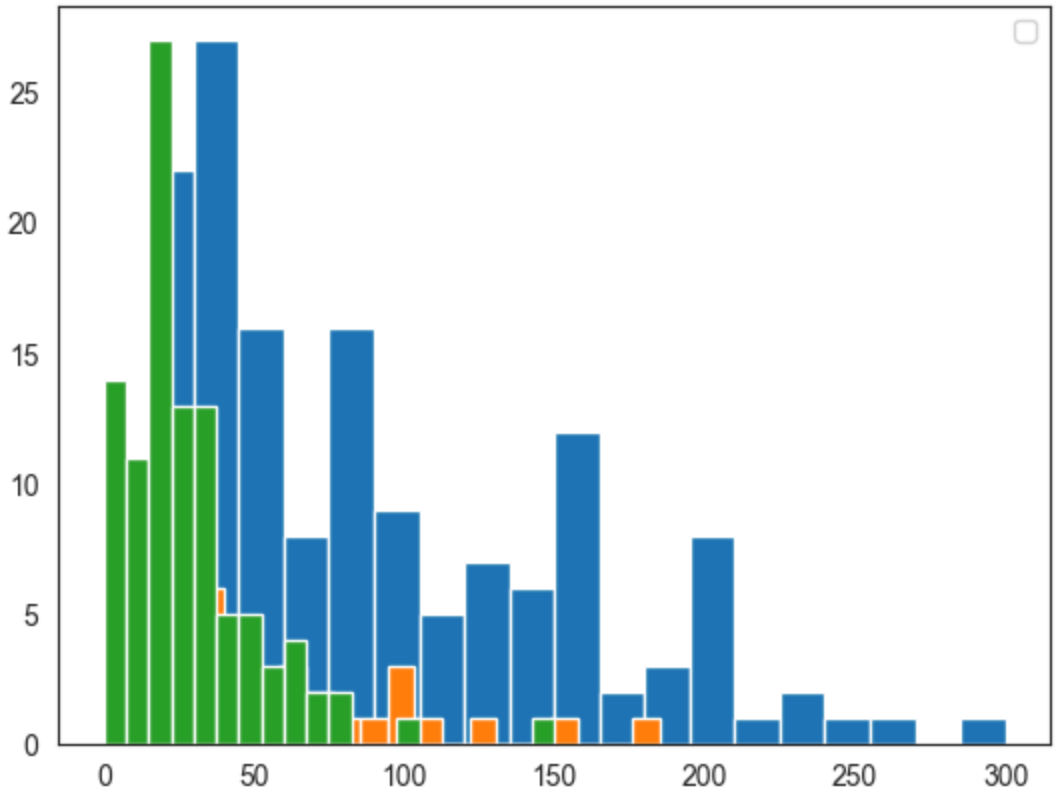
In [84]:
```
movies.head()
```

Out[84]:

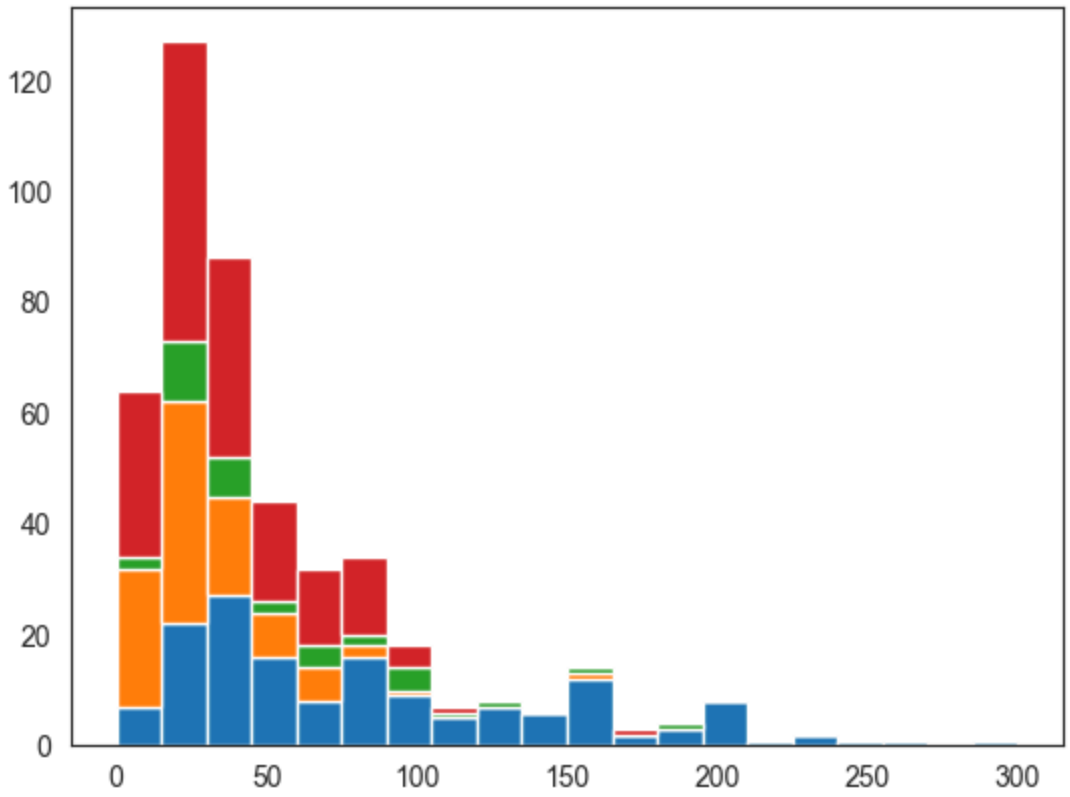
	Film	Genre	CriticRating	AudienceRating	BudgetMillions	Year
0	(500) Days of Summer	Comedy	87	81	8	2009
1	10,000 B.C.	Adventure	9	44	105	2008
2	12 Rounds	Action	30	52	20	2009
3	127 Hours	Adventure	93	84	18	2010
4	17 Again	Comedy	55	70	20	2009

In [85]:
```
#movies.Genre.unique()
```

In [86]:
```
# Below plots are stacked histogram becuase overlaped

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()
```



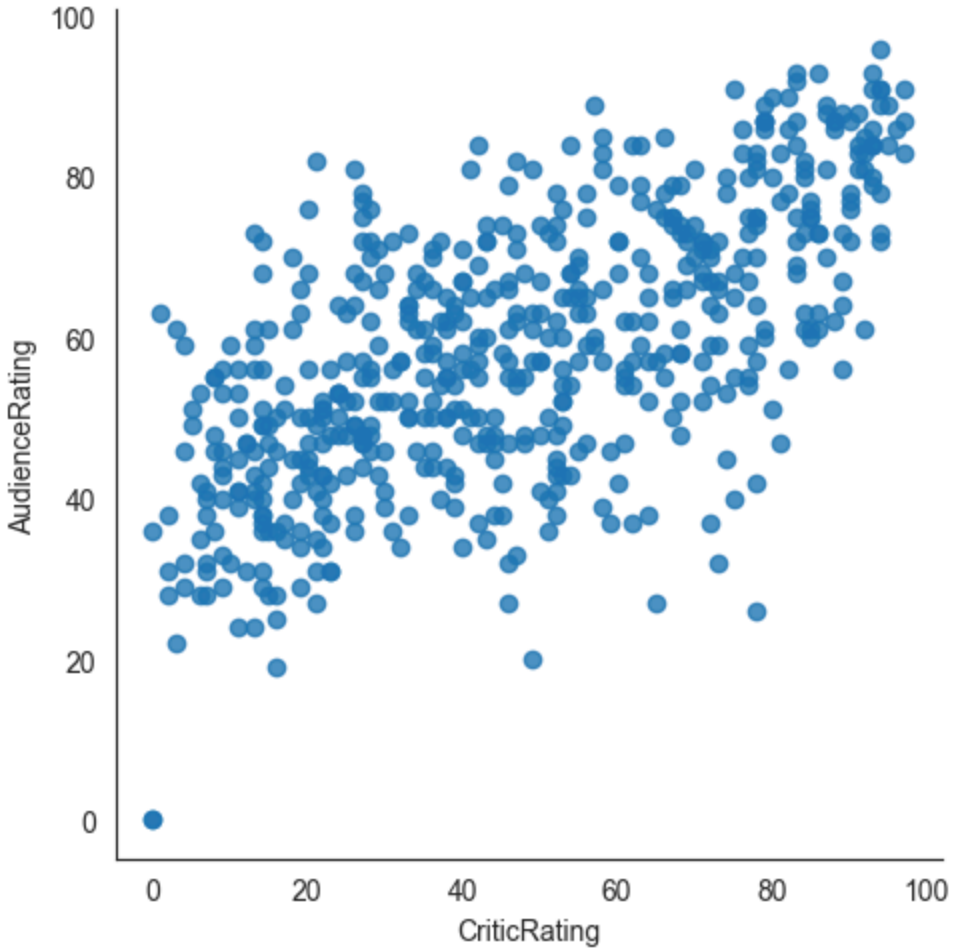
In [87]:
```
plt.hist([movies[movies.Genre == 'Action'].BudgetMillions,\
          movies[movies.Genre == 'Drama'].BudgetMillions, \
          movies[movies.Genre == 'Thriller'].BudgetMillions, \
          movies[movies.Genre == 'Comedy'].BudgetMillions],
         bins = 20, stacked = True)
plt.show()
```



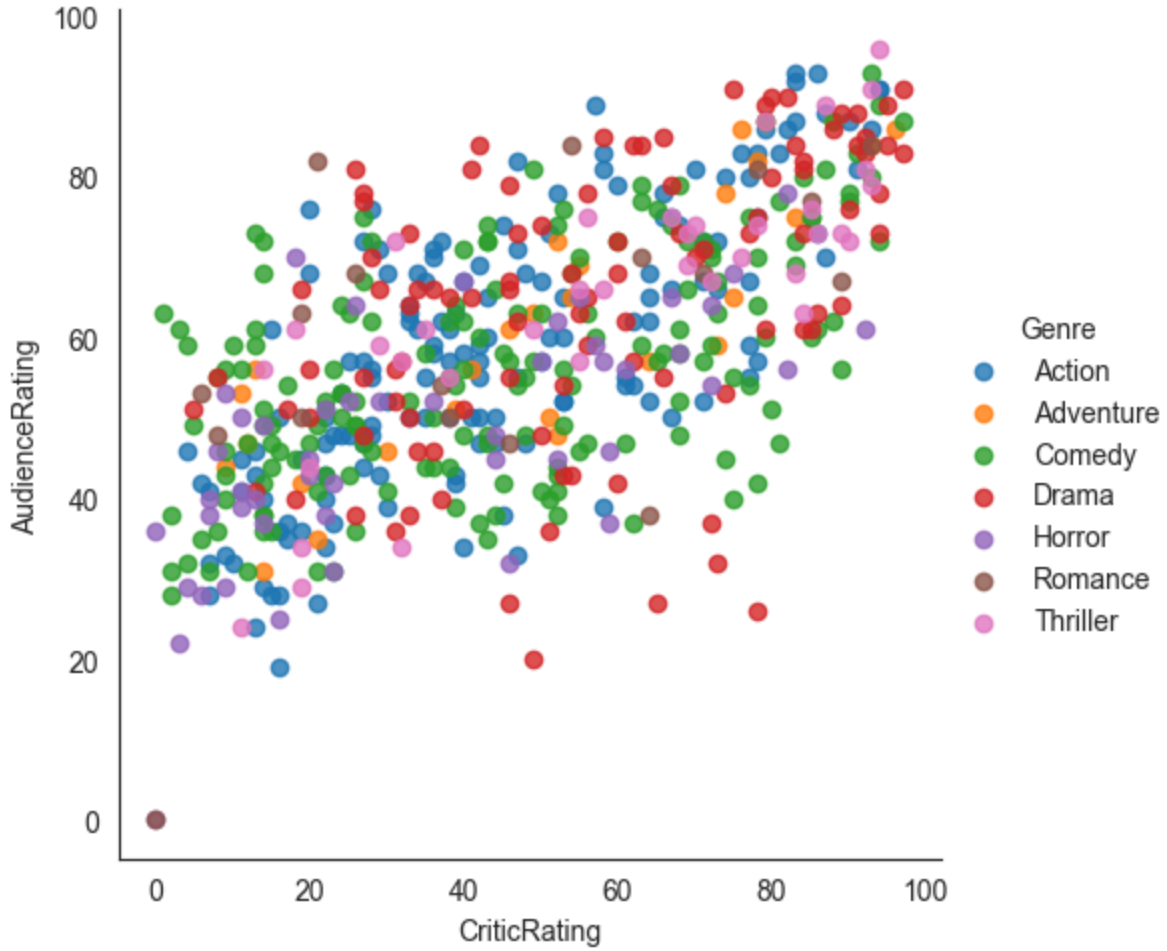
In [88]:
```
# if you have 100 categories you cannot copy & paste all the things

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

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

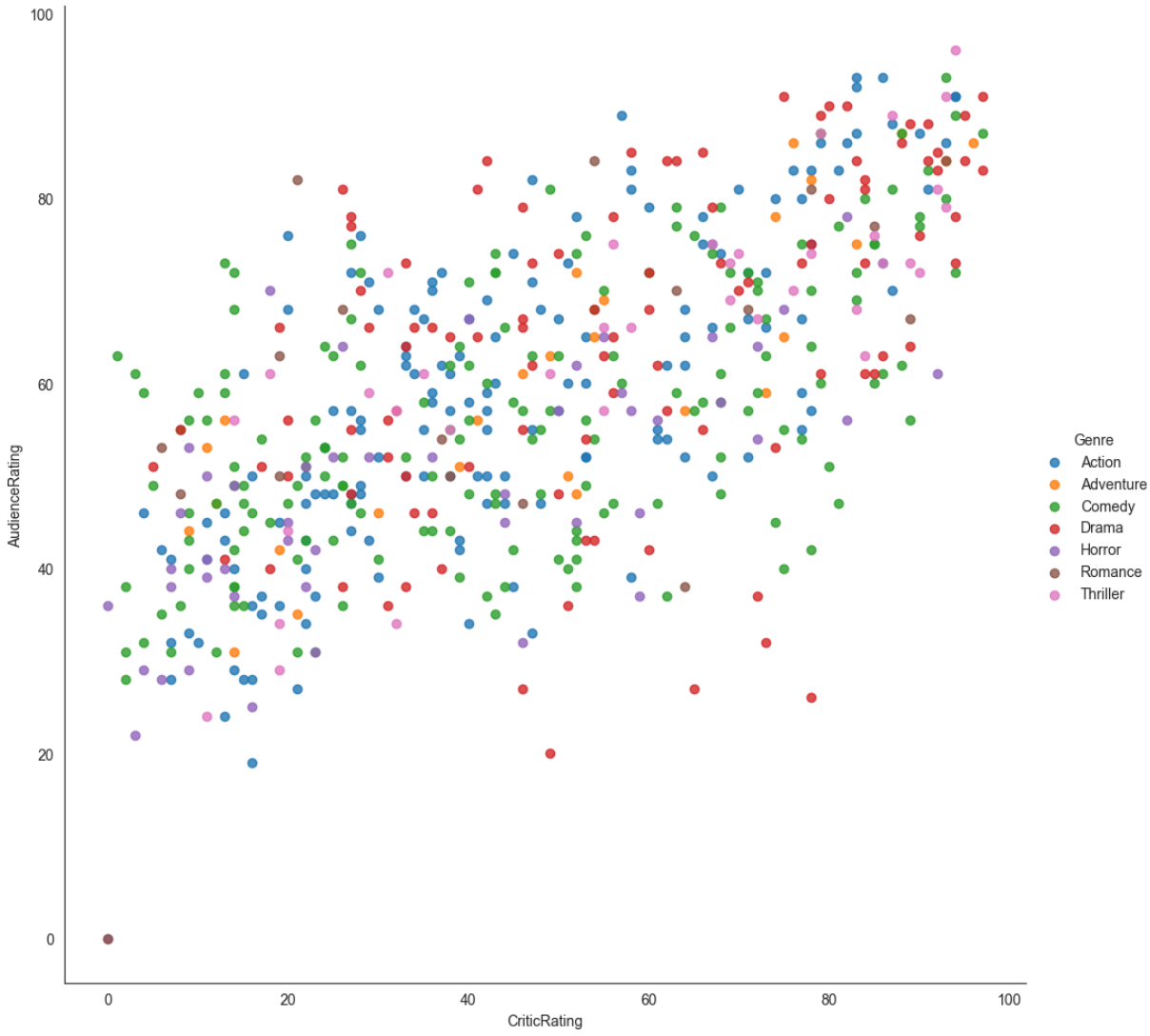
In [89]:
```
vis1 = sns.lmplot(data=movies, x='CriticRating', y='AudienceRating',\
                  fit_reg=False)
```



In [90]:
```
vis1 = sns.lmplot(data=movies, x='CriticRating', y='AudienceRating',\
                  fit_reg=False, hue = 'Genre')
```

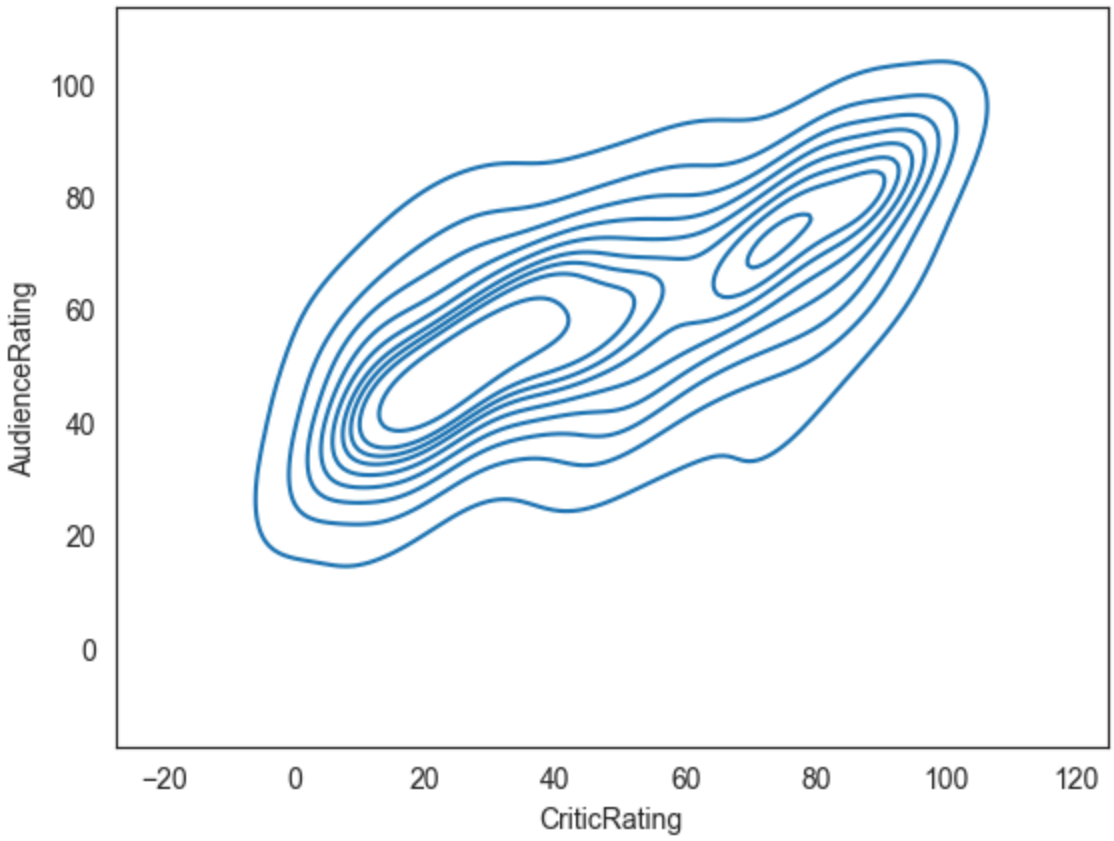


In [96]:
```
vis1 = sns.lmplot(data=movies, x='CriticRating', y='AudienceRating',
                  fit_reg=False, hue = 'Genre', height = 10,aspect=1)
```

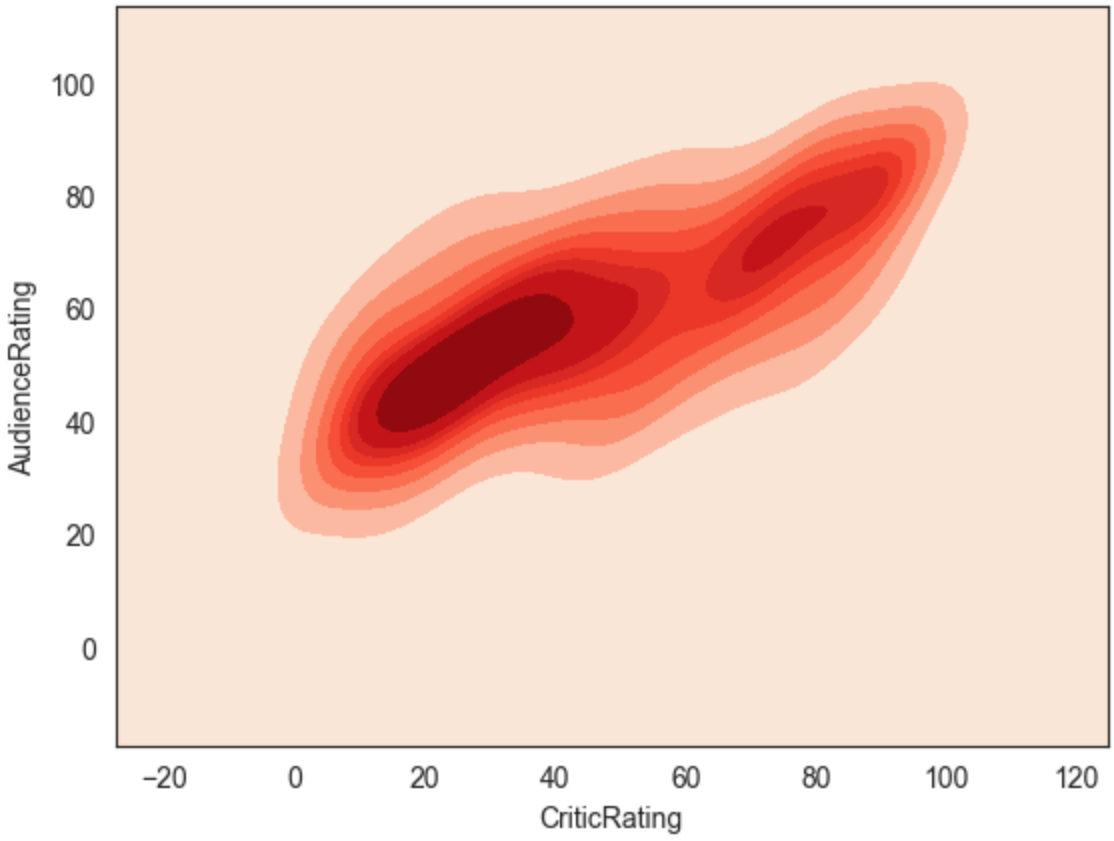


In [109…
```
# Kernal Density Estimate plot ( KDE PLOT)
# how can i visulize audience rating & critics rating . using scatterplot
```

In [113…
```
k1 = sns.kdeplot(data=movies,x="CriticRating",y="AudienceRating")

# where do u find more density and how density is distibuted across from the the ch
# center point is kernal this is calld KDE & insteade of dots it visualize like thi
# we can able to clearly see the spread at the audience ratings
```

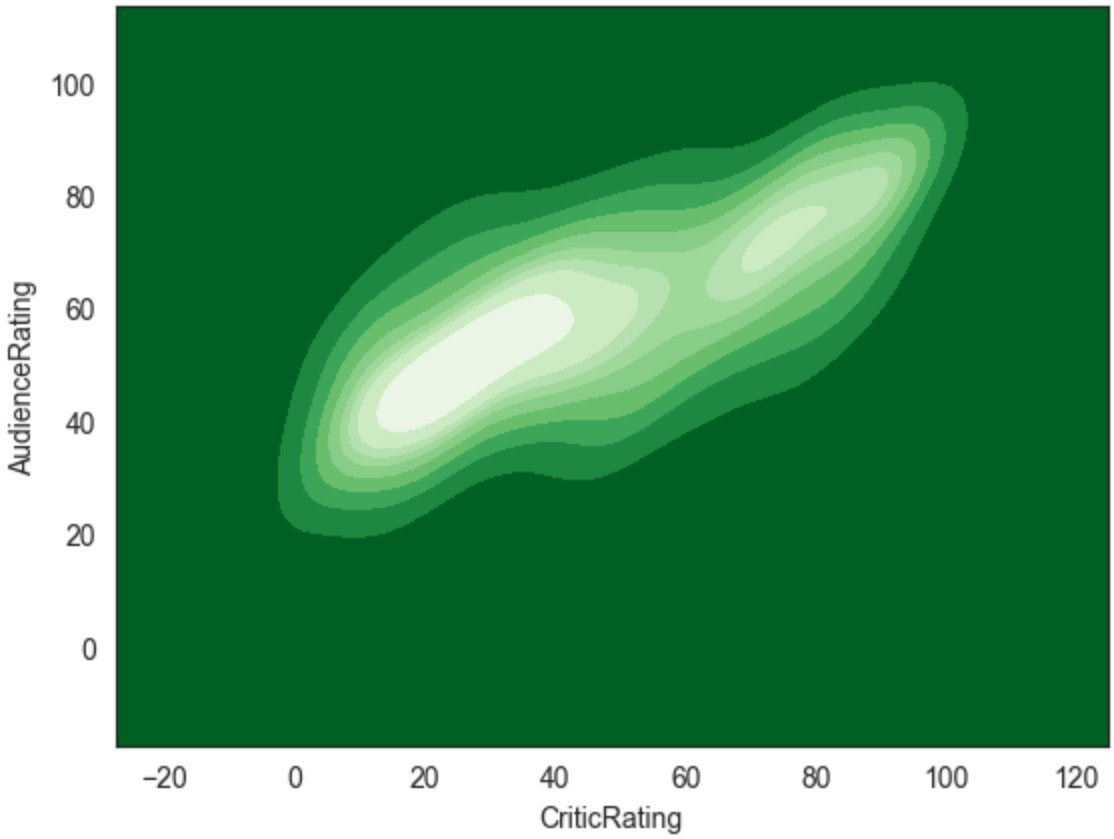


In [116…
```
k1 = sns.kdeplot(data=movies,x="CriticRating",y="AudienceRating",fill=True,cmap="Re
```



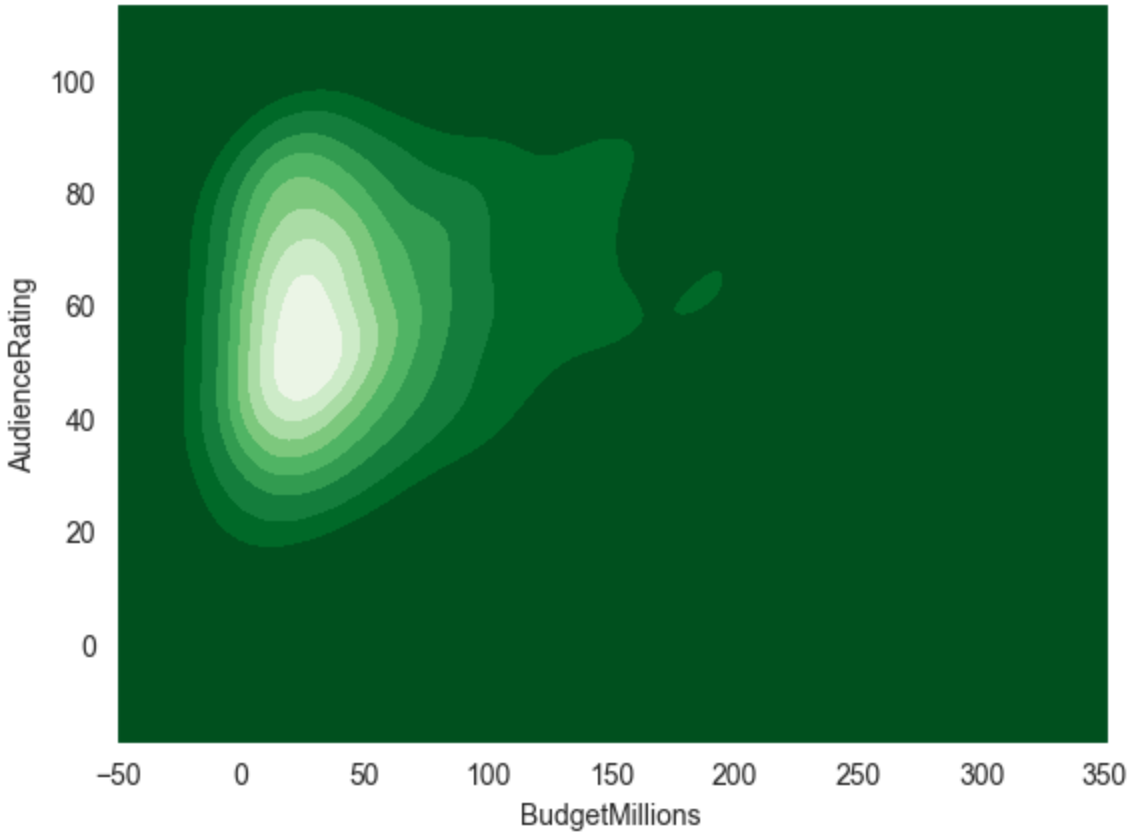
In [118…

```
k2 = sns.kdeplot(
    data=movies,
    x="CriticRating",
    y="AudienceRating",
    fill=True,       # replaces shade
    cmap="Greens_r",
    thresh=0         # similar effect as old shade_lowest=False
)
```



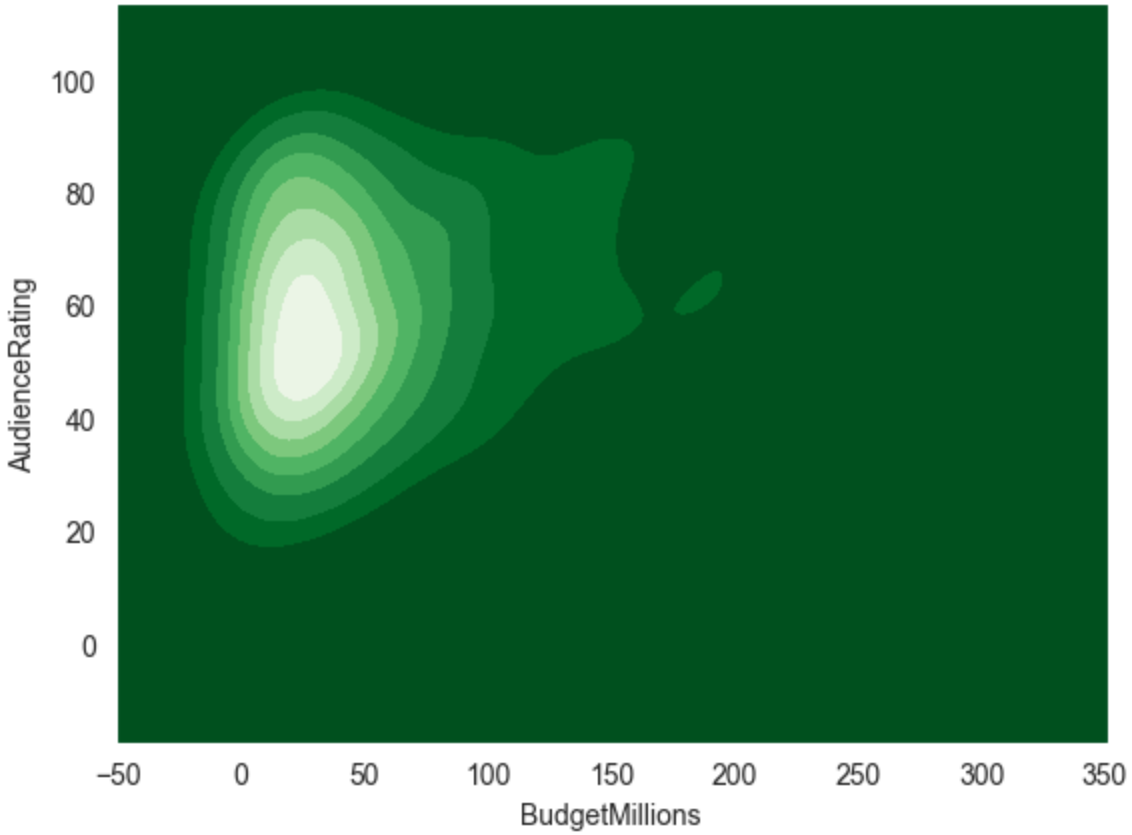
In [121…

```
import seaborn as sns

sns.set_style("dark")

k1 = sns.kdeplot(
    data=movies,
    x="BudgetMillions",
    y="AudienceRating",
    fill=True,        # replaces shade=True
    cmap="Greens_r",
    thresh=0          # replaces shade_lowest=False
)
```



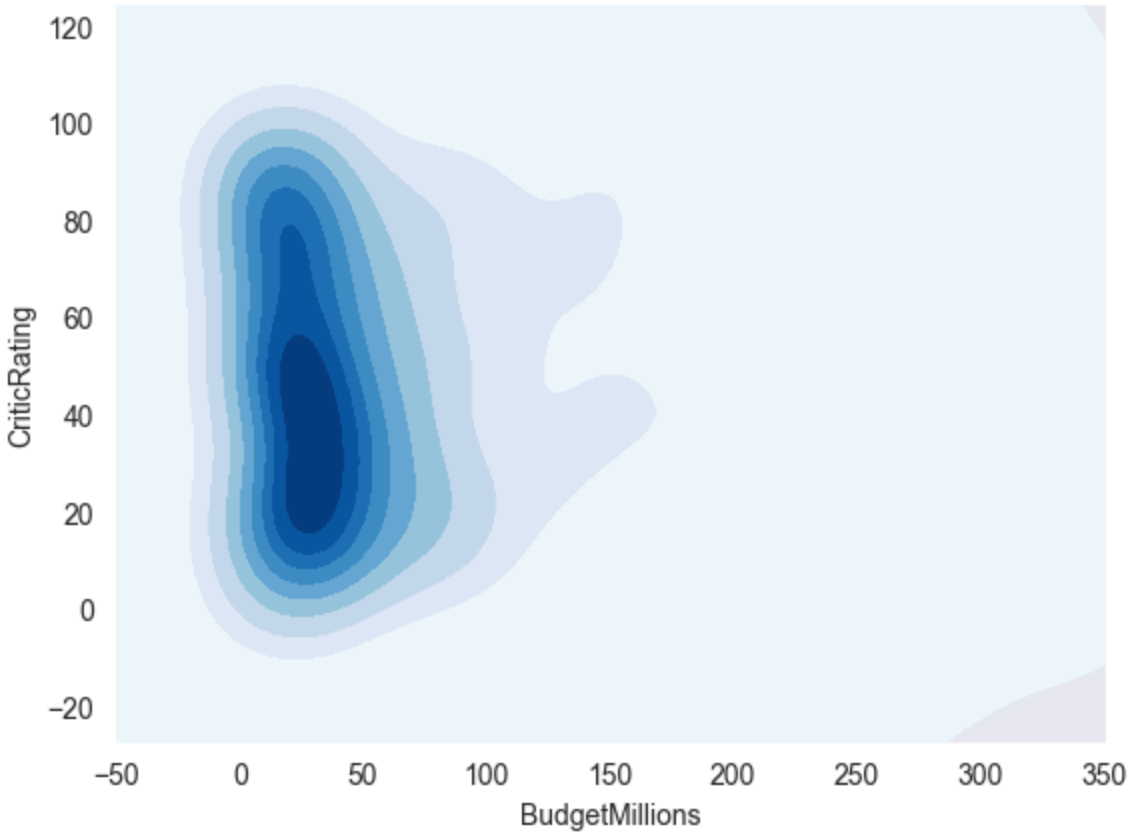
In [125…

```
sns.set_style("dark")

k1 = sns.kdeplot(
    data=movies,
    x="BudgetMillions",
    y="AudienceRating",
    fill=True,       # if you want shaded density
    cmap="Greens_r",
    thresh=0
)
```



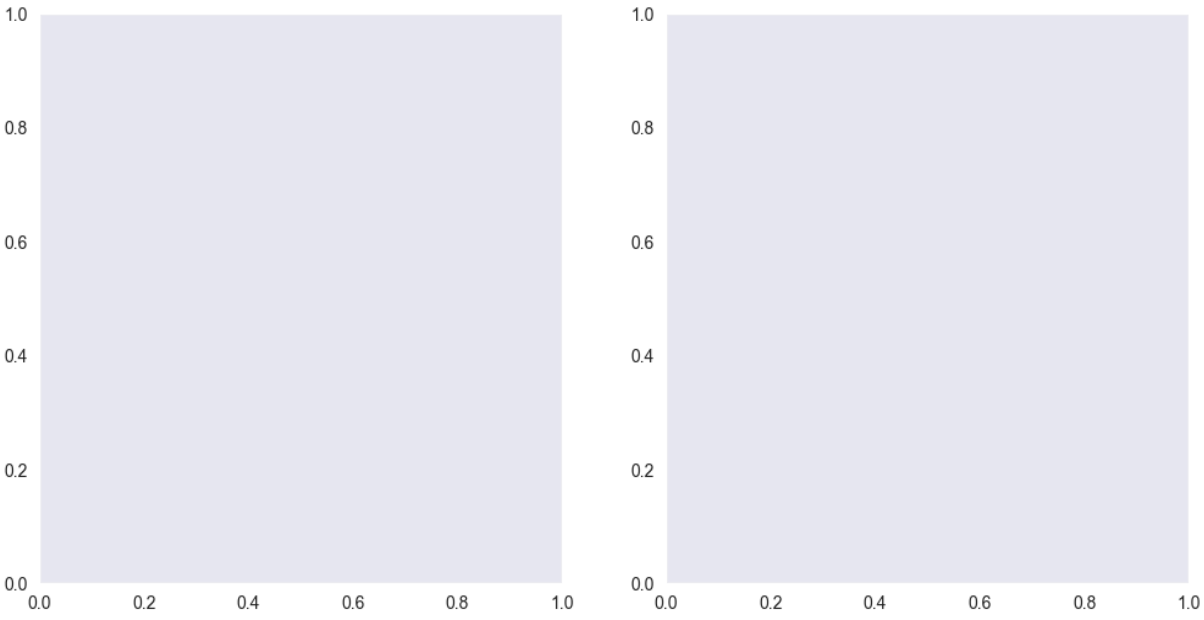
In [127…

```
k2 = sns.kdeplot(
    data=movies,
    x="BudgetMillions",
    y="CriticRating",
    fill=True,       # gives shaded contours
    cmap="Blues",    # you can change color map
    thresh=0
)
```



In [128…

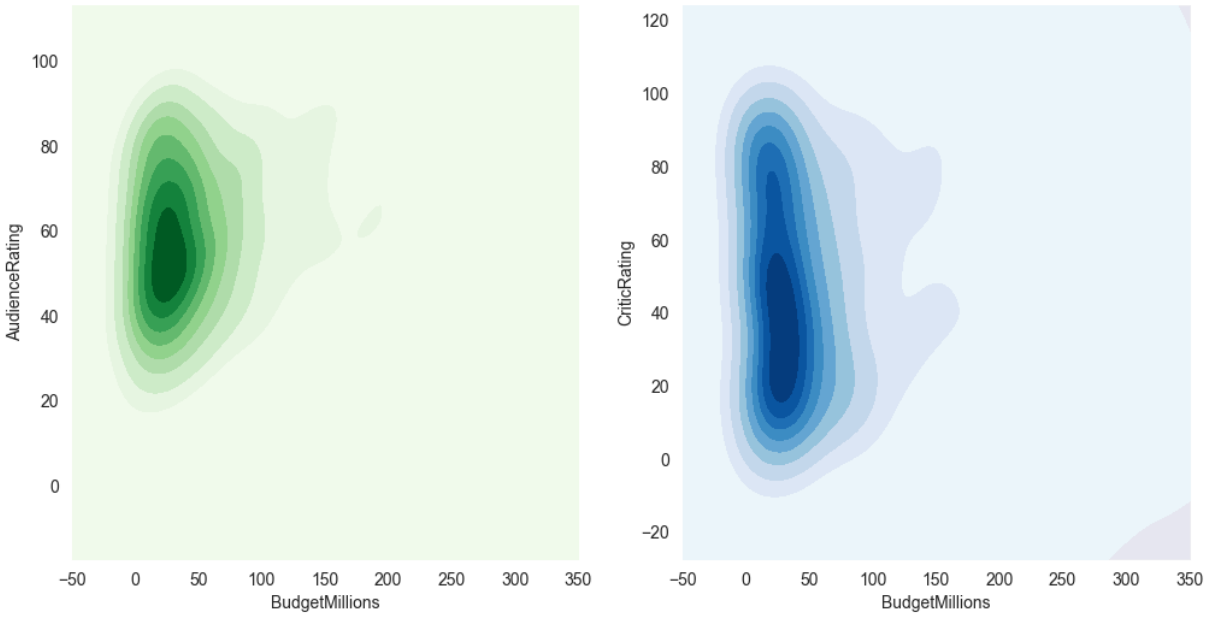
```
#subplots

f, ax = plt.subplots(1,2, figsize =(12,6))
#f, ax = plt.subplots(3,3, figsize =(12,6))
```



In [134…

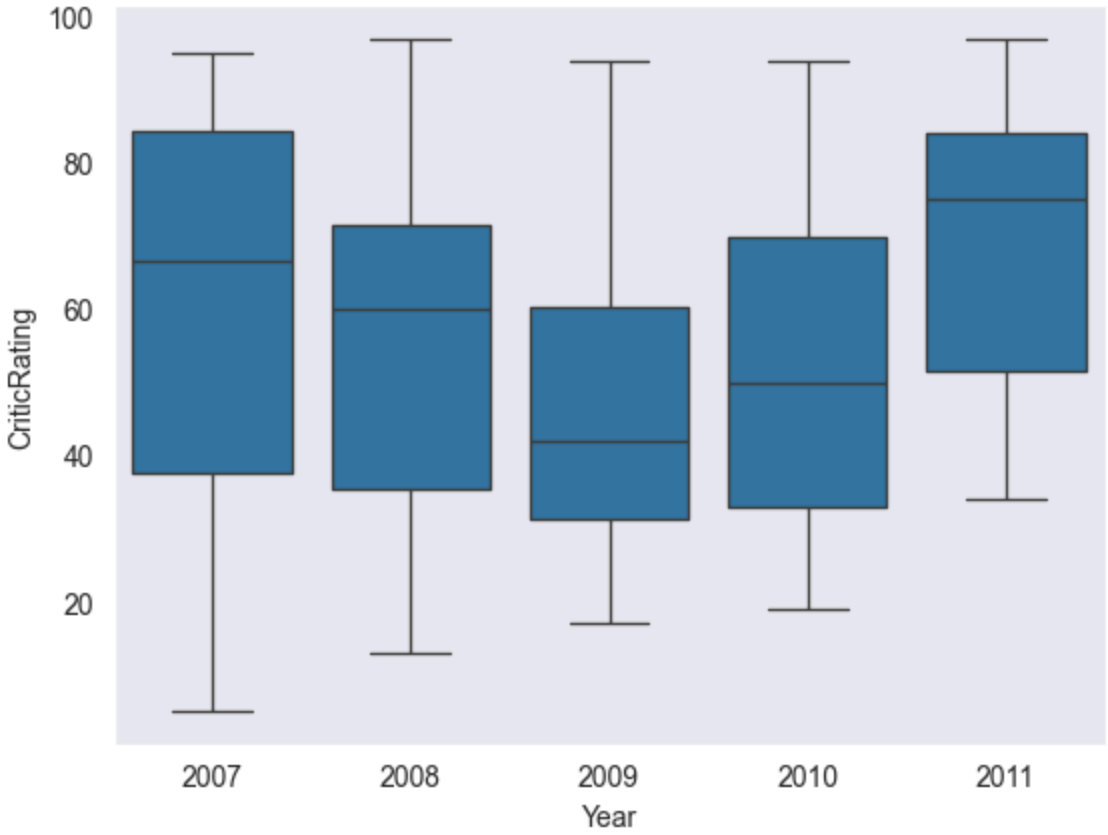
```
import matplotlib.pyplot as plt
import seaborn as sns

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

```
# First KDE plot
k1 = sns.kdeplot(
    data=movies,
    x="BudgetMillions",
    y="AudienceRating",
    fill=True,
    cmap="Greens",
    thresh=0,
    ax=axes[0]
)

# Second KDE plot
k2 = sns.kdeplot(
    data=movies,
    x="BudgetMillions",
    y="CriticRating",
    fill=True,
    cmap="Blues",
    thresh=0,
    ax=axes[1]
)
```



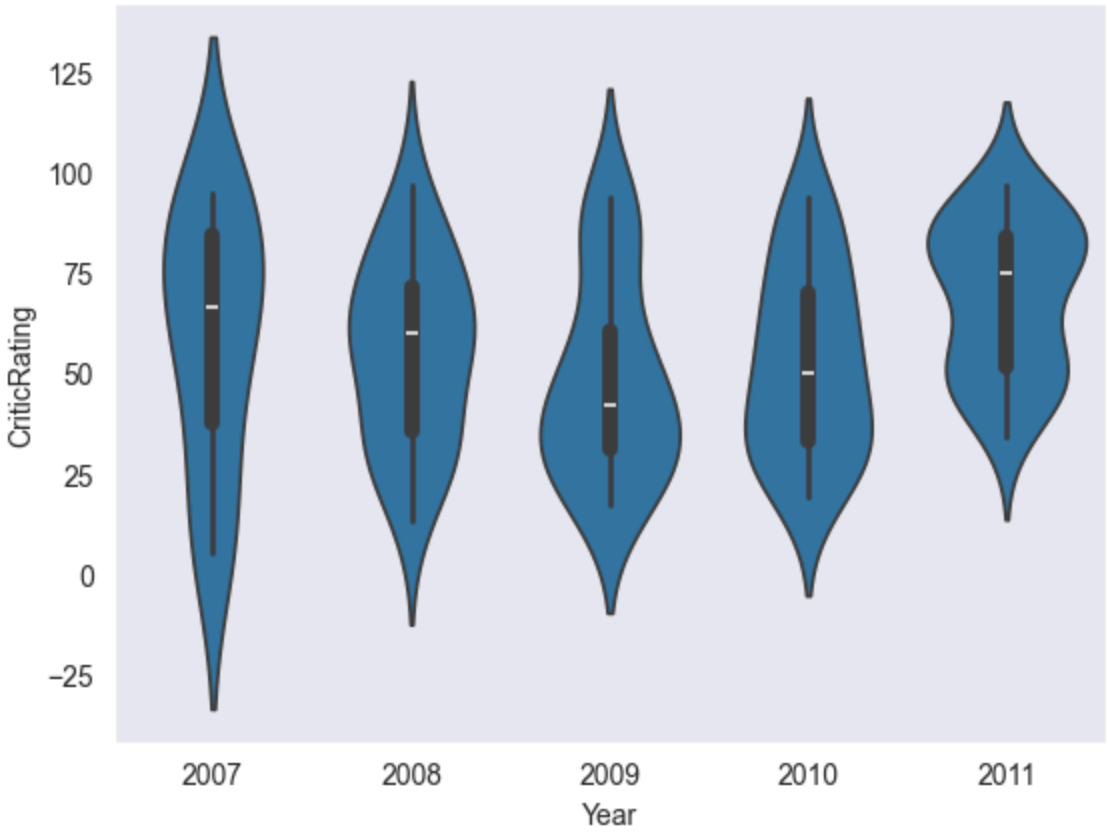
In [137…

```
w1 = sns.boxplot(data=movies[movies.Genre == 'Drama'], x='Year', y = 'CriticRating'
```



In [138…
```
sns.violinplot(data=movies[movies.Genre == 'Drama'], x='Year', y = 'CriticRating')
```

Out[138…
```
<Axes: xlabel='Year', ylabel='CriticRating'>
```



In [140…
```
plt.scatter(movies.CriticRating,movies.AudienceRating)
```

Out[140…
```
<matplotlib.collections.PathCollection at 0x21753702e90>
```