项目：根据TMDb电影数据探索票房的秘密

目录

简介

本报告研究的数据集是TMDb电影数据（已经清洗Kaggle的原始数据），本数据集包含1万条电影信息，信息来源为“电影数据库”（TMDb，The Movie Database），包括用户评分和票房。“演职人员 (cast)”、“电影类别 (genres)”等数据列。

根据数据集所包含的信息，拟研究解决如下几个问题:
a.近几年最受欢迎的电影类型是哪些?
b.票房和电影评分、预算之间的相关性如何?
d.哪些类型的电影可以以低成本获得高票房?
e.哪些公司的电影更能制作出受欢迎的电影?

导入数据包和数据

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
% matplotlib inline
df_movies = pd.read_csv('tmdb-movies.csv')
#设置sns的背景样式
sns.set_style("darkgrid")
```

数据整理

常规属性

In [2]:
```
# 打印数据前几行,检查数据
df_movies.head()
# 打印类型，查看是否有缺失数据或错误数据的情况。
df_movies.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10866 entries, 0 to 10865
Data columns (total 21 columns):
id                      10866 non-null int64
imdb_id                 10856 non-null object
popularity              10866 non-null float64
budget                  10866 non-null int64
revenue                 10866 non-null int64
original_title          10866 non-null object
cast                    10790 non-null object
homepage                2936 non-null object
director                10822 non-null object
tagline                 8042 non-null object
keywords                9373 non-null object
overview                10862 non-null object
runtime                 10866 non-null int64
genres                  10843 non-null object
production_companies    9836 non-null object
release_date            10866 non-null object
vote_count              10866 non-null int64
vote_average            10866 non-null float64
release_year            10866 non-null int64
budget_adj              10866 non-null float64
revenue_adj             10866 non-null float64
dtypes: float64(4), int64(6), object(11)
memory usage: 1.7+ MB
```

数据清理

In [3]:
```
# 去掉一些对分析过程没有帮助的列
df_movies.drop(['budget', 'revenue', 'homepage', 'overview', 'release_date'], axis=1, inplace=True)
# 去掉票房收入和预算为0的行
index_0 = df_movies[df_movies['revenue_adj'] == 0].index.tolist()
df_movies = df_movies.drop(index_0)
index_1 = df_movies[df_movies['budget_adj'] == 0].index.tolist()
df_movies = df_movies.drop(index_1)
```

探索性数据分析

1. 近几年最受欢迎的电影类型是哪些?

In [4]:

```
# 将genres列的类型关键词拆分，并合成新的一列
df_1 = df_movies.drop('genres', axis=1).join(df_movies['genres'].str.split('|', expand=True).s
tack().reset_index(level=1,drop=True).rename('genres'))
df_1.head()
```

Out[4]:

	id	imdb_id	popularity	original_title	cast	director	tagline	
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	monster|dna|tyrar rex|velociraptor|is
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	monster|dna|tyrar rex|velociraptor|is
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	monster|dna|tyrar rex|velociraptor|is
0	135397	tt0369610	32.985763	Jurassic World	Chris Pratt|Bryce Dallas Howard|Irrfan Khan|Vi...	Colin Trevorrow	The park is open.	monster|dna|tyrar rex|velociraptor|is
1	76341	tt1392190	28.419936	Mad Max: Fury Road	Tom Hardy|Charlize Theron|Hugh Keays-Byrne|Nic...	George Miller	What a Lovely Day.	future|chase|post-apocalyptic|dysto

In [5]:
```
#筛选出近十年的电影数据
df_10years = df_1[df_1['release_year'] > 2005]
#将近十年的电影数据按电影类型分组
#分组后将相同类型电影经通货膨胀后的票房累加，从大到小排序
df_10years.groupby('genres').sum()['revenue_adj'].sort_values(ascending=False)
```

Out[5]:
```
genres
Adventure          9.499940e+10
Action             9.188060e+10
Comedy             6.557133e+10
Drama              5.796210e+10
Thriller           5.500410e+10
Fantasy            5.248066e+10
Family             5.127112e+10
Science Fiction    4.787616e+10
Animation          3.573048e+10
Romance            2.448707e+10
Crime              2.338747e+10
Mystery            1.415903e+10
Horror             1.133569e+10
Music              5.507432e+09
War                4.821420e+09
History            3.735904e+09
Western            1.915877e+09
Documentary        5.002484e+08
Foreign            1.107687e+08
Name: revenue_adj, dtype: float64
```

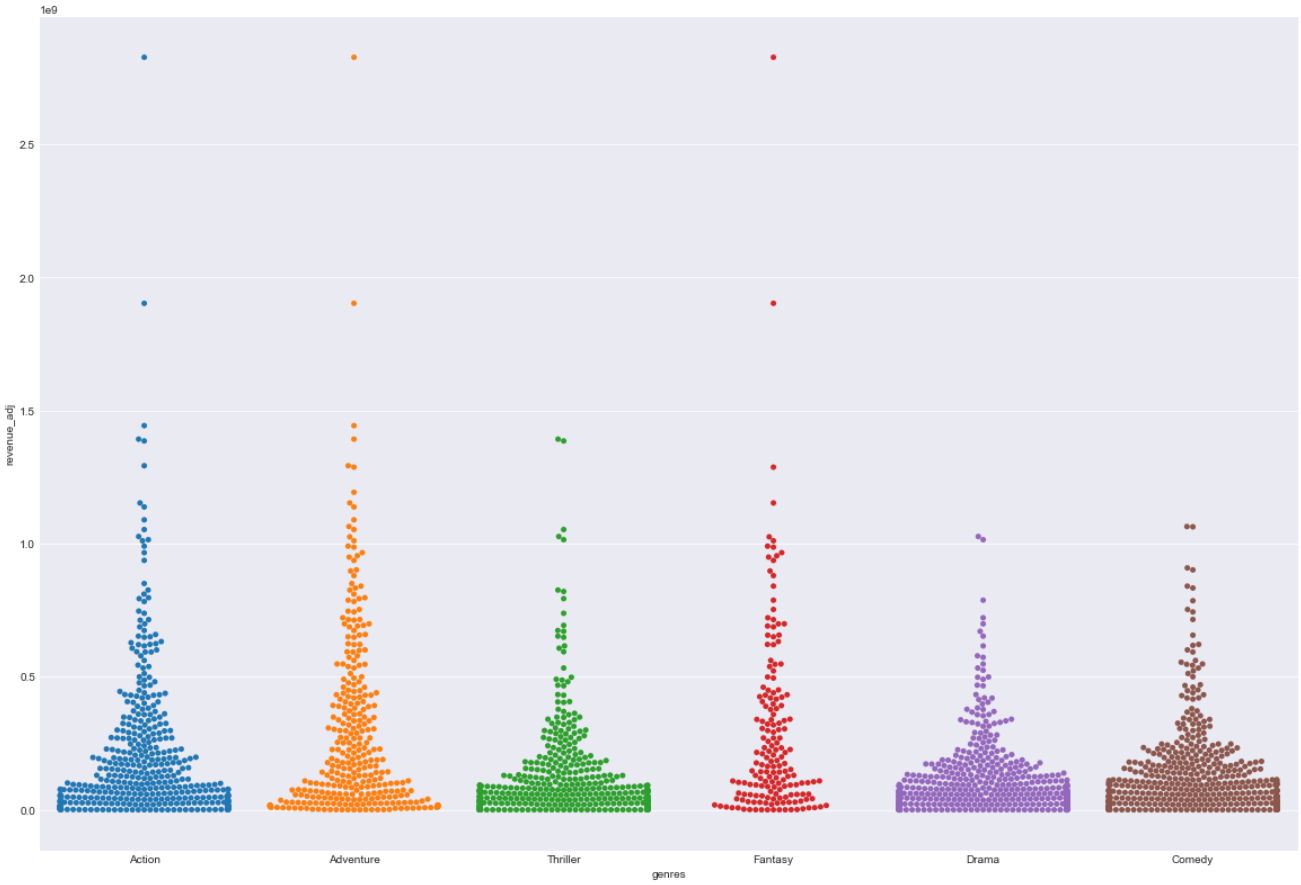
由票房排序结果可以看出近十年来，最受欢迎的电影类型是动作冒险类，其次是喜剧、戏剧以及恐怖类和奇幻类型的电影。

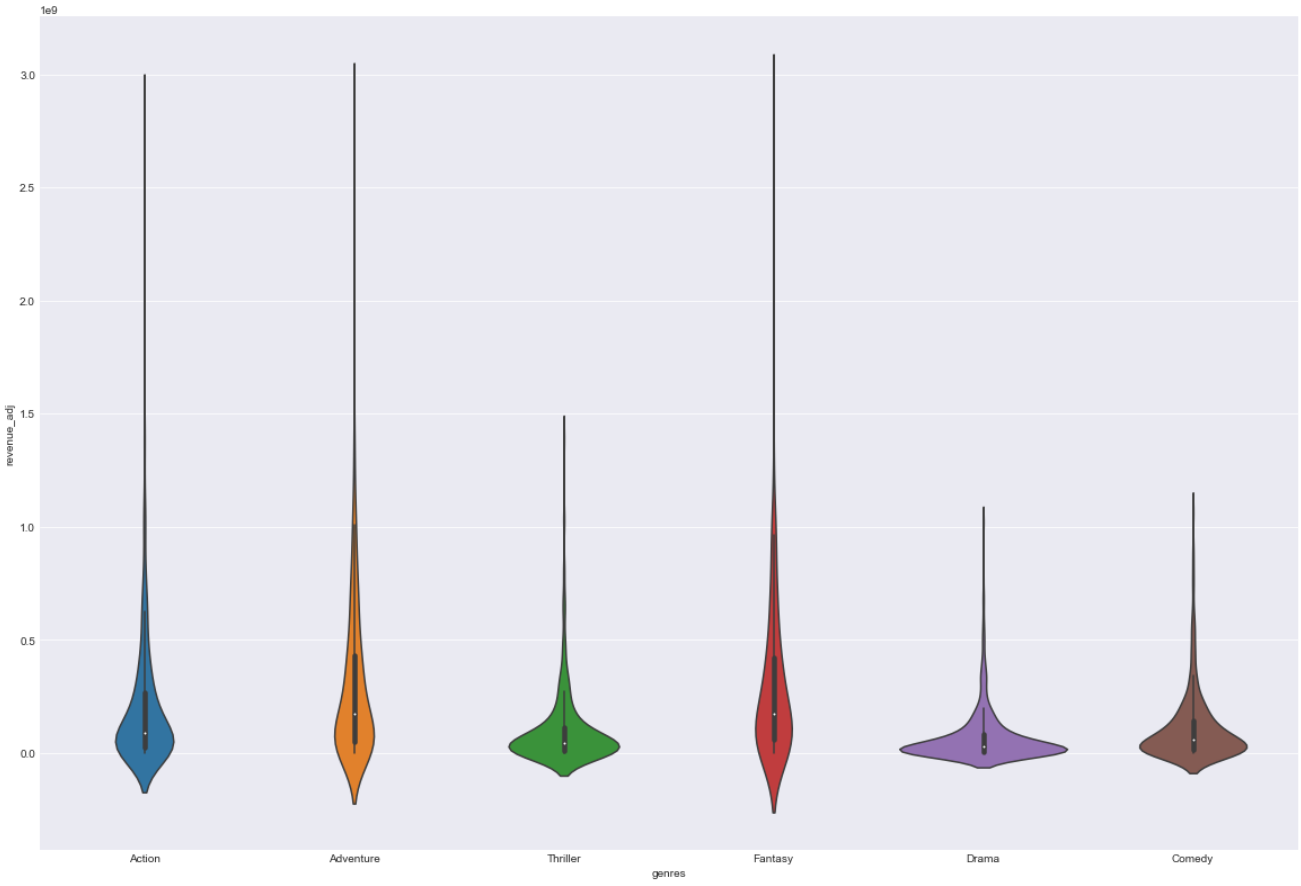
In [6]:
```
#挑选出近年来最受欢迎的几大类型的电影进行进一步分析
category_1 = ['Adventure', 'Fantasy', 'Thriller', 'Drama', 'Action', 'Comedy']
df_10years_6 = df_10years[(df_10years['genres']== 'Adventure')|
                          (df_10years['genres']== 'Fantasy')|
                          (df_10years['genres']== 'Thriller')|
                          (df_10years['genres']== 'Drama')|
                          (df_10years['genres']== 'Action')|
                          (df_10years['genres']== 'Comedy')]
```

In [7]:
```
#对近十年的电影按类型进行分组并对票房求和
df_10years.groupby('genres').sum()['revenue_adj'].sort_values(ascending=False)
```

Out[7]:
```
genres
Adventure          9.499940e+10
Action             9.188060e+10
Comedy             6.557133e+10
Drama              5.796210e+10
Thriller           5.500410e+10
Fantasy            5.248066e+10
Family             5.127112e+10
Science Fiction    4.787616e+10
Animation          3.573048e+10
Romance            2.448707e+10
Crime              2.338747e+10
Mystery            1.415903e+10
Horror             1.133569e+10
Music              5.507432e+09
War                4.821420e+09
History            3.735904e+09
Western            1.915877e+09
Documentary        5.002484e+08
Foreign            1.107687e+08
Name: revenue_adj, dtype: float64
```

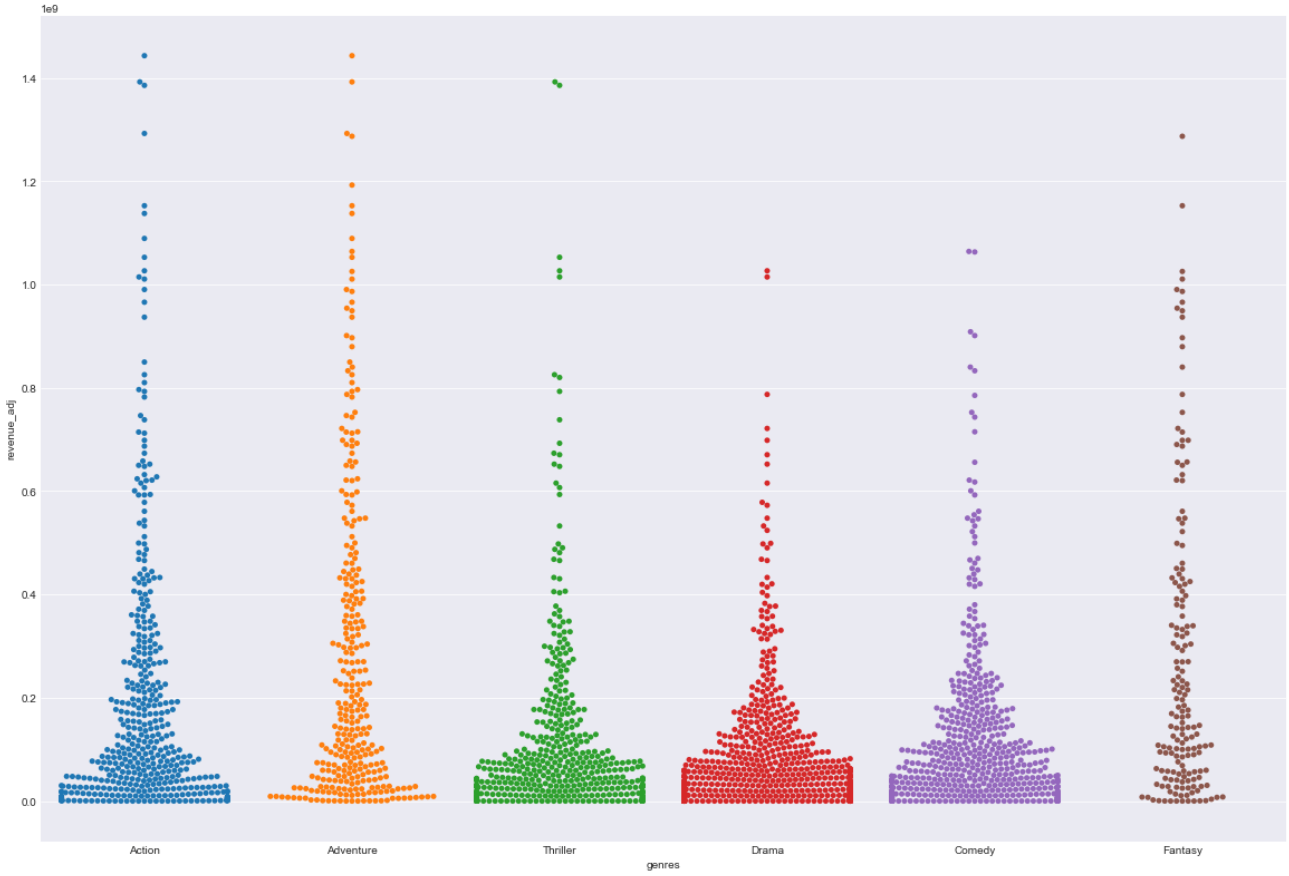
In [8]:
```
#绘制散点图和小提琴图
ax = plt.subplots(figsize=(20,30))
plt.subplot(211)
sns.swarmplot(df_10years_6['genres'], df_10years_6['revenue_adj'],data=df_10years_6)
plt.subplot(212)
sns.violinplot(df_10years_6['genres'], df_10years_6['revenue_adj'],data=df_10years_6);
```

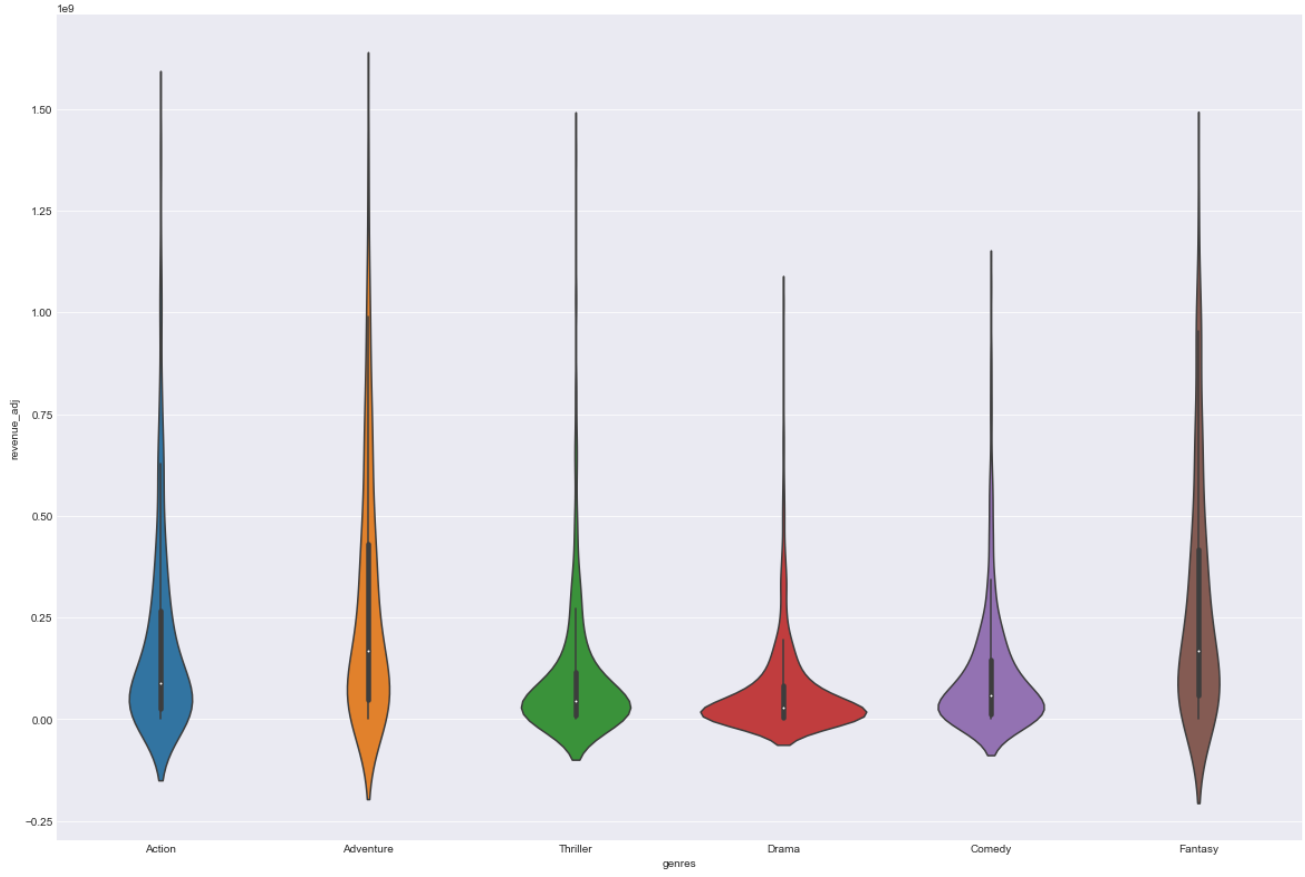




结合散点图和箱线图，可以看到Action、Adventure、Fantasy三种类型的电影有两个相同的较大的异常值，尝试去掉这两个异常值，使图形更便于分析

In [9]:
```
#去掉票房收入大于15亿的两部电影
index_1 = df_10years_6[df_10years_6['revenue_adj'] >= 1.5e+09].index.tolist()
df_10years_6 = df_10years_6.drop(index_1)
#绘制散点图和小提琴图
ax = plt.subplots(figsize=(20,30))
plt.subplot(211)
sns.swarmplot(df_10years_6['genres'], df_10years_6['revenue_adj'],data=df_10years_6)
plt.subplot(212)
sns.violinplot(df_10years_6['genres'], df_10years_6['revenue_adj'],data=df_10years_6);
```

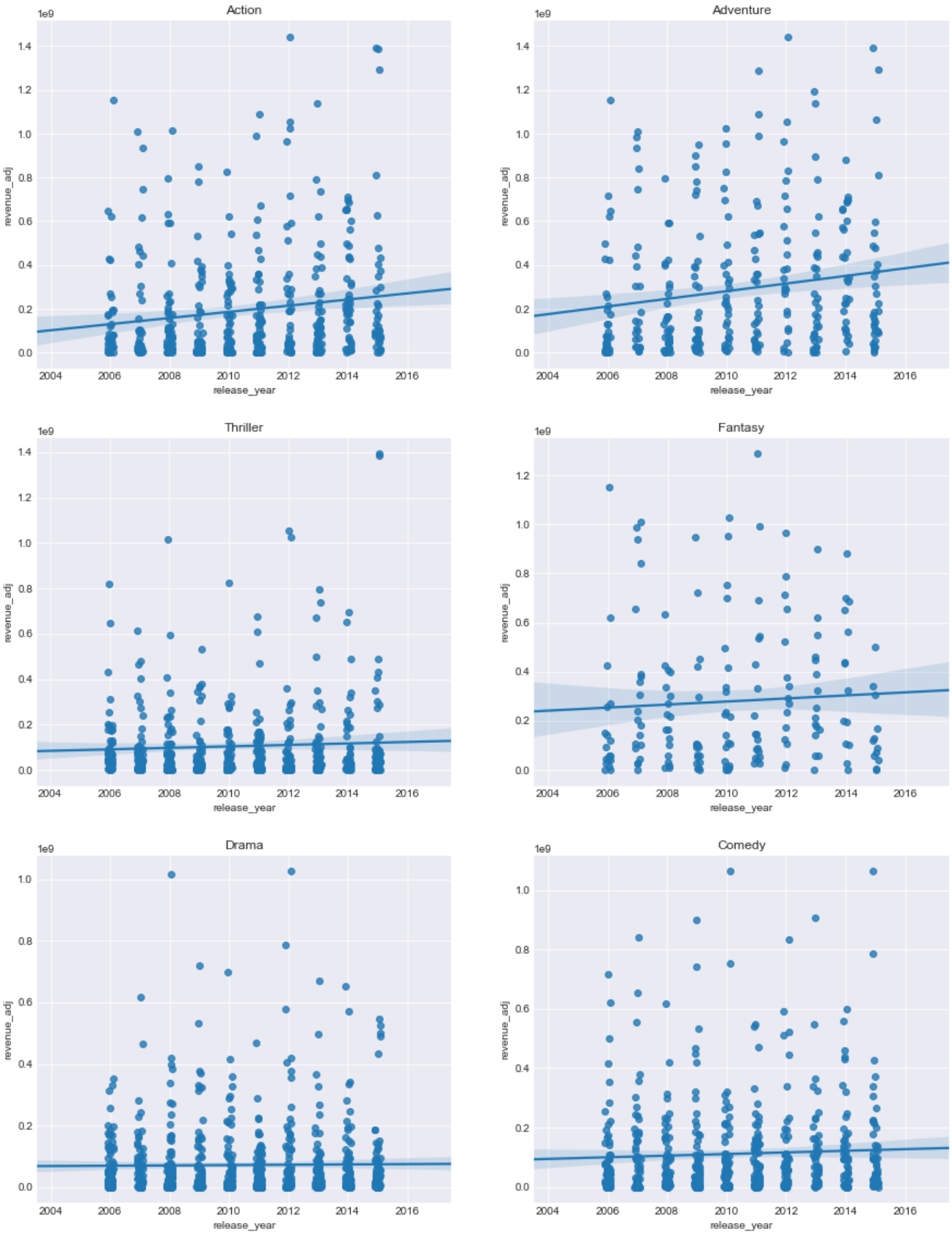




结合散点图和提琴图，整体来看，六大类电影的票房分布都有较高的右偏态。而且近十年内恐怖片、喜剧片以及戏剧类电影数量较多，但票房都集中在在2亿以内，动作冒险类以及科幻类的电影更有潜力拍出票房收益可观的作品。

接下来探索随时间变化，哪些类型的电影越来越受欢迎。

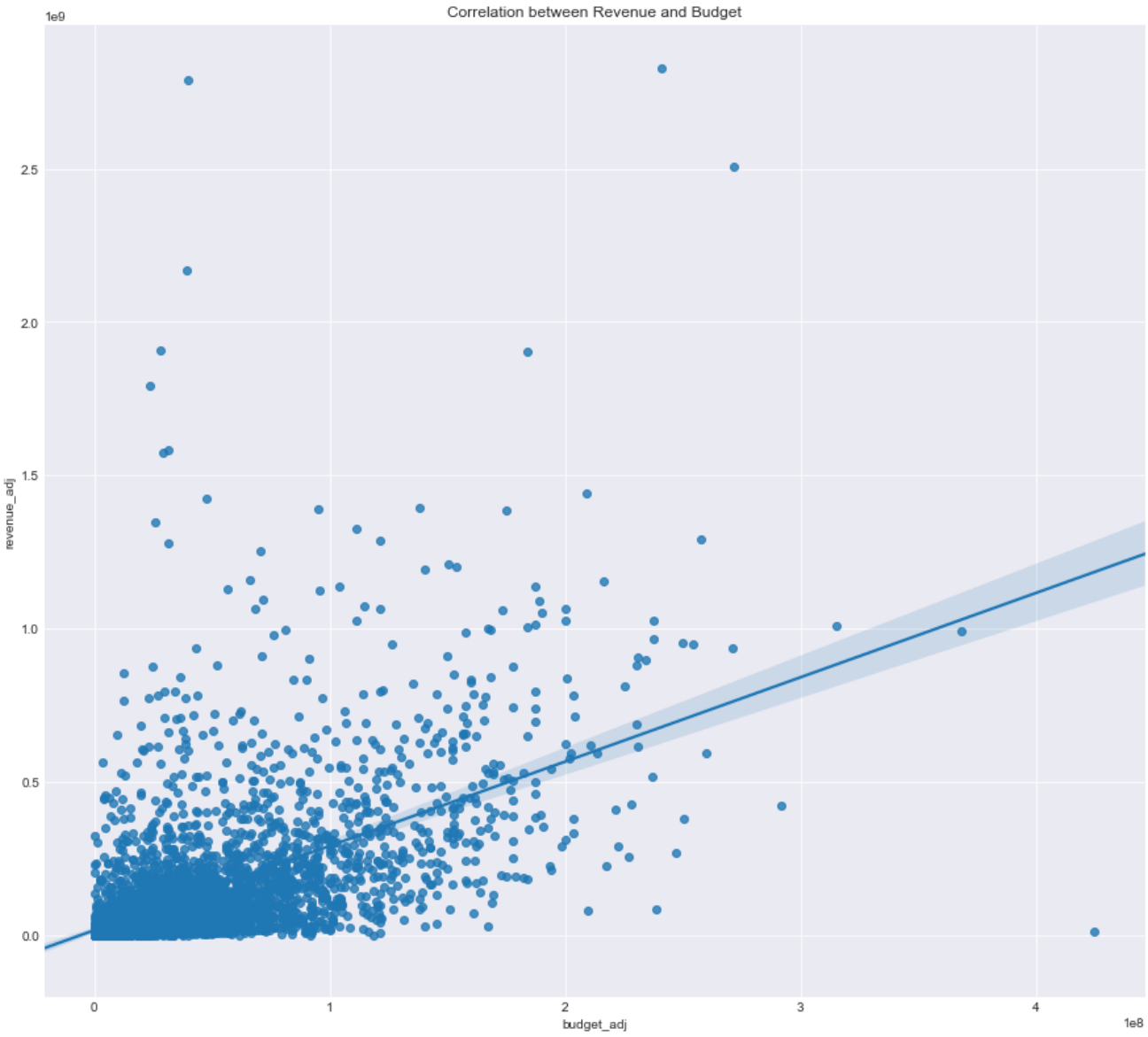
In [10]:
```
#分别绘制六大类型电影随时间变化的票房收入的散点图
ax = plt.subplots(figsize=(15,20))
plt.subplot(321)
plt.title('Action')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Action'], x_jitter=.1)
plt.subplot(322)
plt.title('Adventure')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Adventure'], x_jitter=.1)
plt.subplot(323)
plt.title('Thriller')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Thriller'], x_jitter=.1)
plt.subplot(324)
plt.title('Fantasy')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Fantasy'], x_jitter=.1)
plt.subplot(325)
plt.title('Drama')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Drama'], x_jitter=.1)
plt.subplot(326)
plt.title('Comedy')
sns.regplot(x='release_year', y='revenue_adj', data=df_10years_6[df_10years_6['genres'] == 'Comedy'], x_jitter=.1);
```

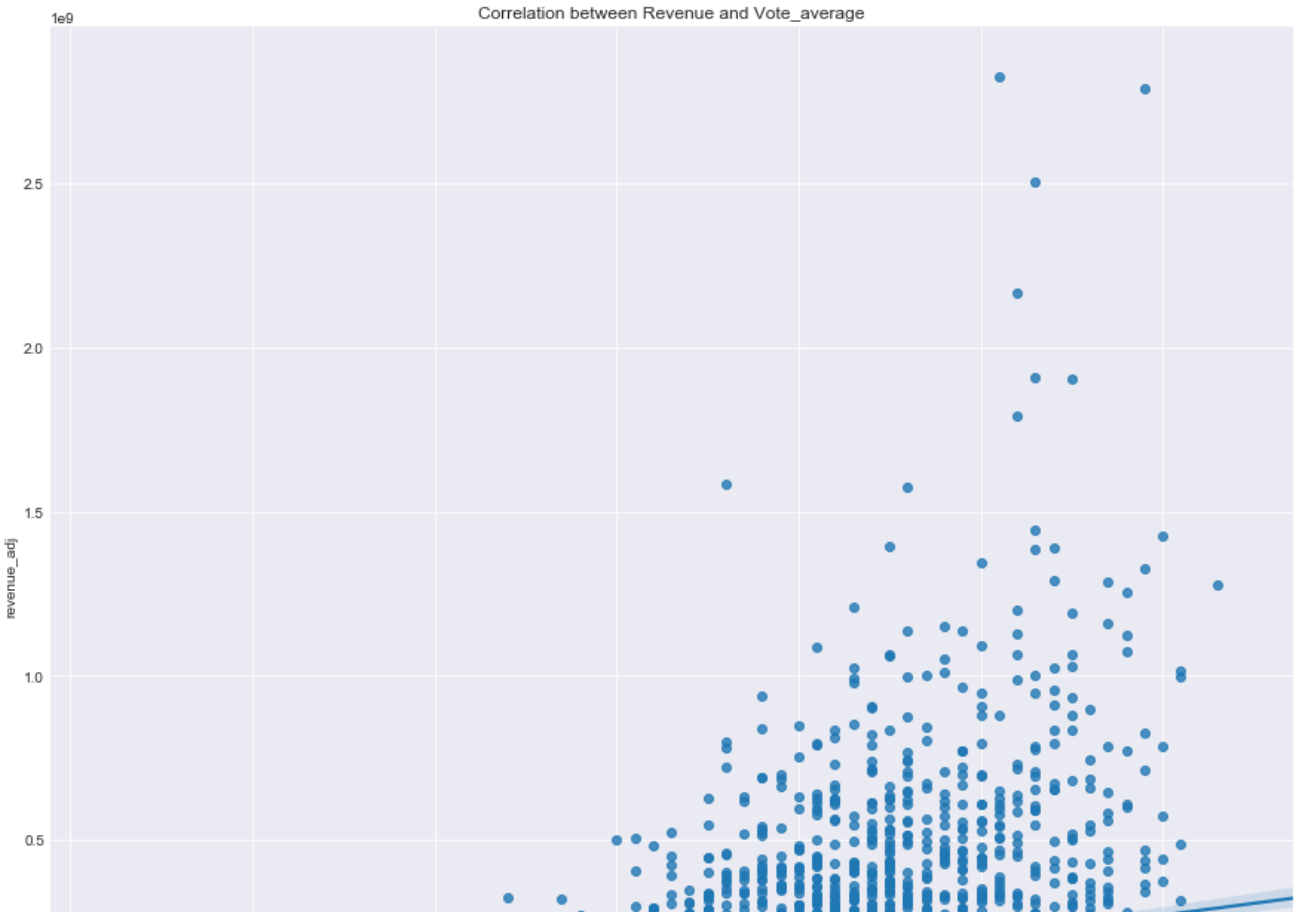


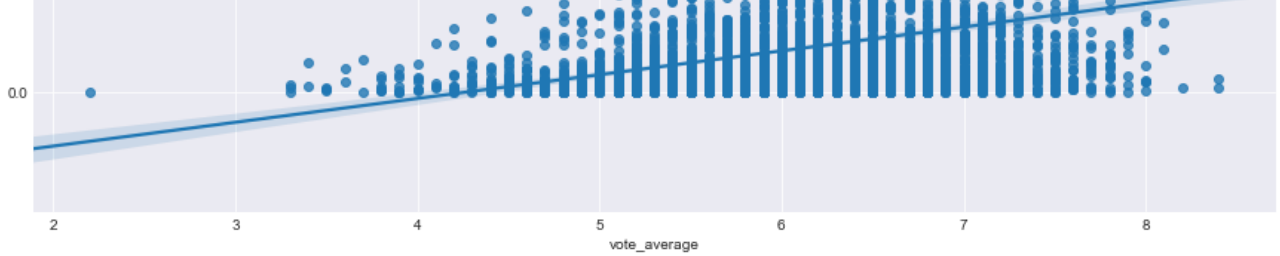
由散点图的拟合回归线可以看出，只有动作类和冒险类的电影和发行时间有较为明显的正相关关系，其余几大类型的电影没有显示出和时间比较明显的相关关系。

2. 票房和电影评分、预算之间的相关性如何?

In [11]:
```
#分别绘制票房收入和电影平均评分以及电影预算的散点图
df_2 = df_movies
ax = plt.subplots(figsize=(15,30))
plt.subplot(211)
plt.title('Correlation between Revenue and Budget')
sns.regplot(x='budget_adj', y='revenue_adj', data=df_2)
plt.subplot(212)
plt.title('Correlation between Revenue and Vote_average')
sns.regplot(x='vote_average', y='revenue_adj', data=df_2);
```


Correlation between Revenue and Budget
1e9
2.5
2.0
1.5
1.0
0.5
0.0
revenue_adj
0
1
2
3
4
budget_adj
1e8


Correlation between Revenue and Vote_average
1e9
2.5
2.0
1.5
1.0
0.5
revenue_adj



由散点图和回归线，电影的票房收入和预算有明显的正相关关系，而票房和电影平均评分有一定的正相关关系但是相关关系不强。

3. 近几年哪些类型的电影可以以低成本获得高票房?

In [19]:
```
#以电影票房收入和电影预算的比值添加新的一列
df_3 = df_10years.copy()
df_3['rate'] = df_3['revenue_adj']/df_3['budget_adj']
#观察票房收入和预算比值的大概范围
df_3.sort_values(by='rate')
```

Out[19]:

	id	imdb_id	popularity	original_title	cast	director	tagline
6707	9986	tt0413895	0.536631	Charlotte's Web	Julia Roberts|Steve Buscemi|John Cleese|Oprah ...	Gary Winick	Something unexpected, unbelievable, unforgetta...
6707	9986	tt0413895	0.536631	Charlotte's Web	Julia Roberts|Steve Buscemi|John Cleese|Oprah ...	Gary Winick	Something unexpected, unbelievable, unforgetta...
6707	9986	tt0413895	0.536631	Charlotte's Web	Julia Roberts|Steve Buscemi|John Cleese|Oprah ...	Gary Winick	Something unexpected, unbelievable, unforgetta...
7506	2196	tt0795368	0.642207	Death at a Funeral	Matthew Macfadyen|Alan Tudyk|Peter Dinklage|Ke...	Frank Oz	From director Frank Oz comes the story of a fa...
7506	2196	tt0795368	0.642207	Death at a Funeral	Matthew Macfadyen|Alan Tudyk|Peter Dinklage|Ke...	Frank Oz	From director Frank Oz comes the story of a fa...
3239	14301	tt1227926	0.352054	Dr. Horrible's Sing-Along Blog	Neil Patrick Harris|Nathan Fillion|Felicia Day...	Joss Whedon	He has a Ph.D. in horribleness!
3239	14301	tt1227926	0.352054	Dr. Horrible's Sing-Along Blog	Neil Patrick Harris|Nathan Fillion|Felicia Day...	Joss Whedon	He has a Ph.D. in horribleness!
3239	14301	tt1227926	0.352054	Dr. Horrible's Sing-Along Blog	Neil Patrick Harris|Nathan Fillion|Felicia Day...	Joss Whedon	He has a Ph.D. in horribleness!
3239	14301	tt1227926	0.352054	Dr. Horrible's Sing-Along Blog	Neil Patrick Harris|Nathan Fillion|Felicia Day...	Joss Whedon	He has a Ph.D. in horribleness!
3239	14301	tt1227926	0.352054	Dr. Horrible's Sing-Along Blog	Neil Patrick Harris|Nathan Fillion|Felicia Day...	Joss Whedon	He has a Ph.D. in horribleness!
7602	13910	tt0472071	0.405537	Death Defying Acts	Catherine Zeta-Jones|Guy Pearce|Timothy Spall|...	Gillian Armstrong	There is no escape.
7602	13910	tt0472071	0.405537	Death Defying Acts	Catherine Zeta-Jones|Guy Pearce|Timothy Spall|...	Gillian Armstrong	There is no escape.

	id	imdb_id	popularity	original_title	cast	director	tagline
4611	98339	tt1867093	0.402086	The Samaritan	Samuel L. Jackson|Luke Kirby|Ruth Negga|Tom Wi...	David Weaver	NaN
5586	227707	tt1376213	0.744520	The Adventurer: The Curse of the Midas Box	Aneurin Barnard|Michael Sheen|Lena Headey|Sam ...	Jonathan Newman	The new name for adventure.
5586	227707	tt1376213	0.744520	The Adventurer: The Curse of the Midas Box	Aneurin Barnard|Michael Sheen|Lena Headey|Sam ...	Jonathan Newman	The new name for adventure.
5586	227707	tt1376213	0.744520	The Adventurer: The Curse of the Midas Box	Aneurin Barnard|Michael Sheen|Lena Headey|Sam ...	Jonathan Newman	The new name for adventure.
3752	65650	tt1582271	0.244803	The Good Doctor	Orlando Bloom|Riley Keough|Taraji P. Henson|Ro...	Lance Daly	Do no harm.
3752	65650	tt1582271	0.244803	The Good Doctor	Orlando Bloom|Riley Keough|Taraji P. Henson|Ro...	Lance Daly	Do no harm.
3573	50601	tt1486193	0.539108	5 Days of War	Rupert Friend|Val Kilmer|Andy GarcÃ-a|Dean Cai...	Renny Harlin	Their only weapon is the truth.
3573	50601	tt1486193	0.539108	5 Days of War	Rupert Friend|Val Kilmer|Andy GarcÃ-a|Dean Cai...	Renny Harlin	Their only weapon is the truth.
5704	158916	tt2224004	0.387592	Sweetwater	Ed Harris|January Jones|Jason Isaacs|Eduardo N...	Logan Miller	Revenge is Sweet.
5704	158916	tt2224004	0.387592	Sweetwater	Ed Harris|January Jones|Jason Isaacs|Eduardo N...	Logan Miller	Revenge is Sweet.
4591	89691	tt1603257	0.436617	ATM	Alice Eve|Josh Peck|Brian Geraghty|Aaron Hughe...	David Brooks	No warning. No control. No escape.
4591	89691	tt1603257	0.436617	ATM	Alice Eve|Josh Peck|Brian Geraghty|Aaron Hughe...	David Brooks	No warning. No control. No escape.

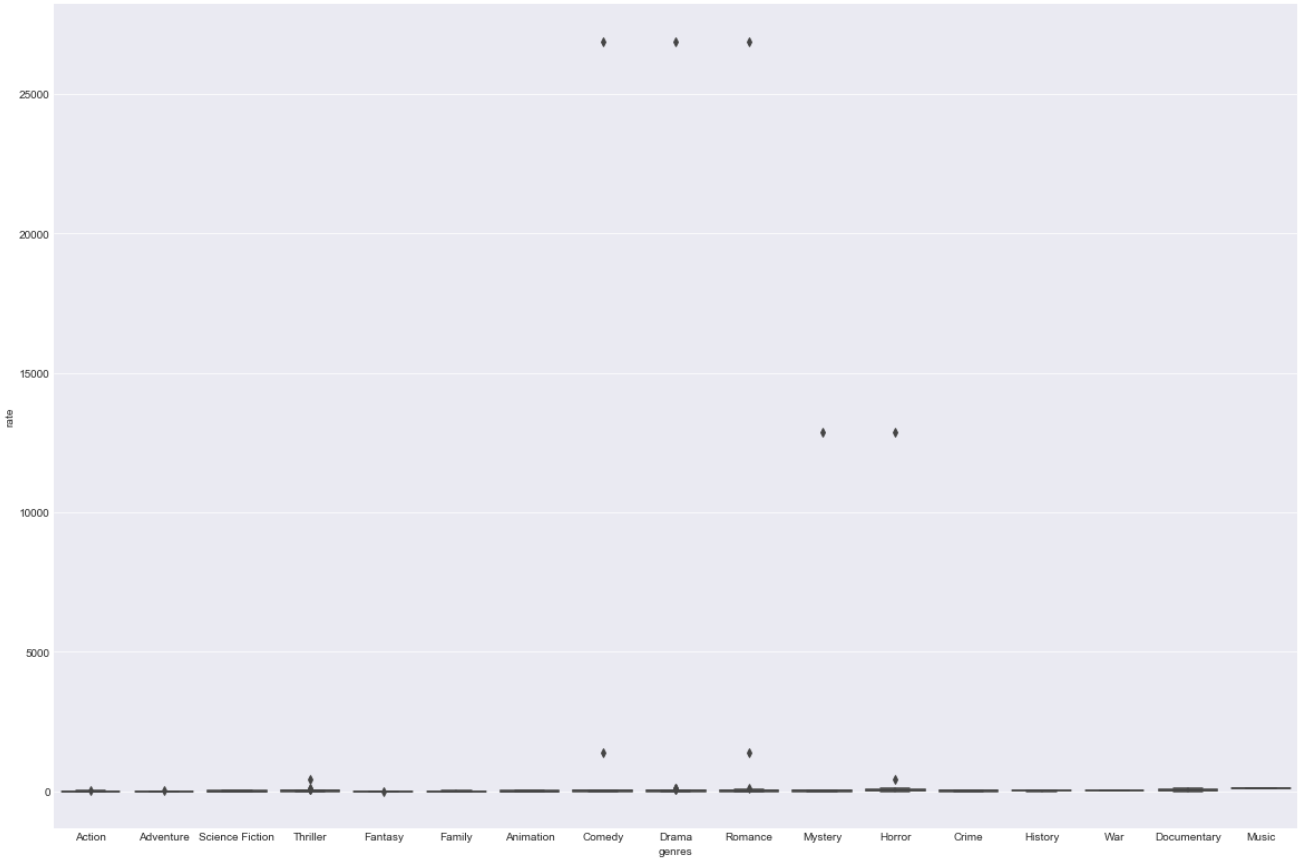
	id	imdb_id	popularity	original_title	cast	director	tagline
4859	116977	tt0249516	0.111351	Foodfight!	Charlie Sheen|Wayne Brady|Hilary Duff|Eva Long...	Lawrence Kasanoff	When good food... goes bad!
4859	116977	tt0249516	0.111351	Foodfight!	Charlie Sheen|Wayne Brady|Hilary Duff|Eva Long...	Lawrence Kasanoff	When good food... goes bad!
4859	116977	tt0249516	0.111351	Foodfight!	Charlie Sheen|Wayne Brady|Hilary Duff|Eva Long...	Lawrence Kasanoff	When good food... goes bad!
4859	116977	tt0249516	0.111351	Foodfight!	Charlie Sheen|Wayne Brady|Hilary Duff|Eva Long...	Lawrence Kasanoff	When good food... goes bad!
5576	93828	tt1684233	0.794369	Welcome to the Punch	James McAvoy|Mark Strong|David Morrissey|Peter...	Eran Creevy	A Stunning, Intelligent Thriller
5576	93828	tt1684233	0.794369	Welcome to the Punch	James McAvoy|Mark Strong|David Morrissey|Peter...	Eran Creevy	A Stunning, Intelligent Thriller
...	...	...	...	...	...	...	...
2044	38358	tt1320244	0.761889	The Last Exorcism	Ashley Bell|Patrick Fabian|Iris Bahr|Louis Her...	Daniel Stamm	Believe In Him.
762	250546	tt3322940	1.018115	Annabelle	Annabelle Wallis|Alfre Woodard|Eric Ladin|Tony...	John R. Leonetti	Before the Conjuring, there was Annabelle.
3523	72571	tt1778304	0.760193	Paranormal Activity 3	Katie Featherston|Sprague Grayden|Lauren Bittn...	Henry Joost|Ariel Schulman	It Runs In The Family
6838	1781	tt0497116	0.251388	An Inconvenient Truth	Al Gore|Billy West|Ronald Reagan|George W. Bus...	Davis Guggenheim	By far the most terrifying film you will ever ...
3755	79120	tt1714210	0.243777	Weekend	Tom Cullen|Chris New|Jonathan Race|Laura Freem...	Andrew Haigh	A (sort of) love story between two guys over a...

	id	imdb_id	popularity	original_title	cast	director	tagline
3755	79120	tt1714210	0.243777	Weekend	Tom Cullen|Chris New|Jonathan Race|Laura Freem...	Andrew Haigh	A (sort of) love story between two guys over a...
2022	41436	tt1536044	0.875432	Paranormal Activity 2	Katie Featherston|David Bierend|Brian Boland|M...	Tod Williams	In 2009 you demanded it. Nothing can prepare y...
2022	41436	tt1536044	0.875432	Paranormal Activity 2	Katie Featherston|David Bierend|Brian Boland|M...	Tod Williams	In 2009 you demanded it. Nothing can prepare y...
136	277685	tt3713166	1.191138	Unfriended	Shelley Hennig|Moses Jacob Storm|Renee Olstead...	Levan Gabriadze	Online, your memories last forever. But so do ...
136	277685	tt3713166	1.191138	Unfriended	Shelley Hennig|Moses Jacob Storm|Renee Olstead...	Levan Gabriadze	Online, your memories last forever. But so do ...
1964	49018	tt1591095	1.396973	Insidious	Patrick Wilson|Rose Byrne|Barbara Hershey|Leig...	James Wan	It's not the house that's haunted.
1964	49018	tt1591095	1.396973	Insidious	Patrick Wilson|Rose Byrne|Barbara Hershey|Leig...	James Wan	It's not the house that's haunted.
3180	14438	tt1129423	0.280555	Fireproof	Kirk Cameron|Erin Bethea|Ken Bevel|Stephen Der...	Alex Kendrick	Never Leave Your Partner Behind.
837	193612	tt2235779	0.667732	The Quiet Ones	Jared Harris|Sam Claflin|Olivia Cooke|Erin Ric...	John Pogue	A shocking experiment. An unspeakable evil.
2207	42296	tt1584016	0.346071	Catfish	Megan Faccio|Melody C. Roscher|Ariel Schulman|...	Henry Joost|Ariel Schulman	Don't let anyone tell you what it is.
4560	76487	tt1560985	0.491648	The Devil Inside	Fernanda Andrade|Simon Quarterman|Evan Helmuth...	William Brent Bell	No soul is safe.
4560	76487	tt1560985	0.491648	The Devil Inside	Fernanda Andrade|Simon Quarterman|Evan Helmuth...	William Brent Bell	No soul is safe.

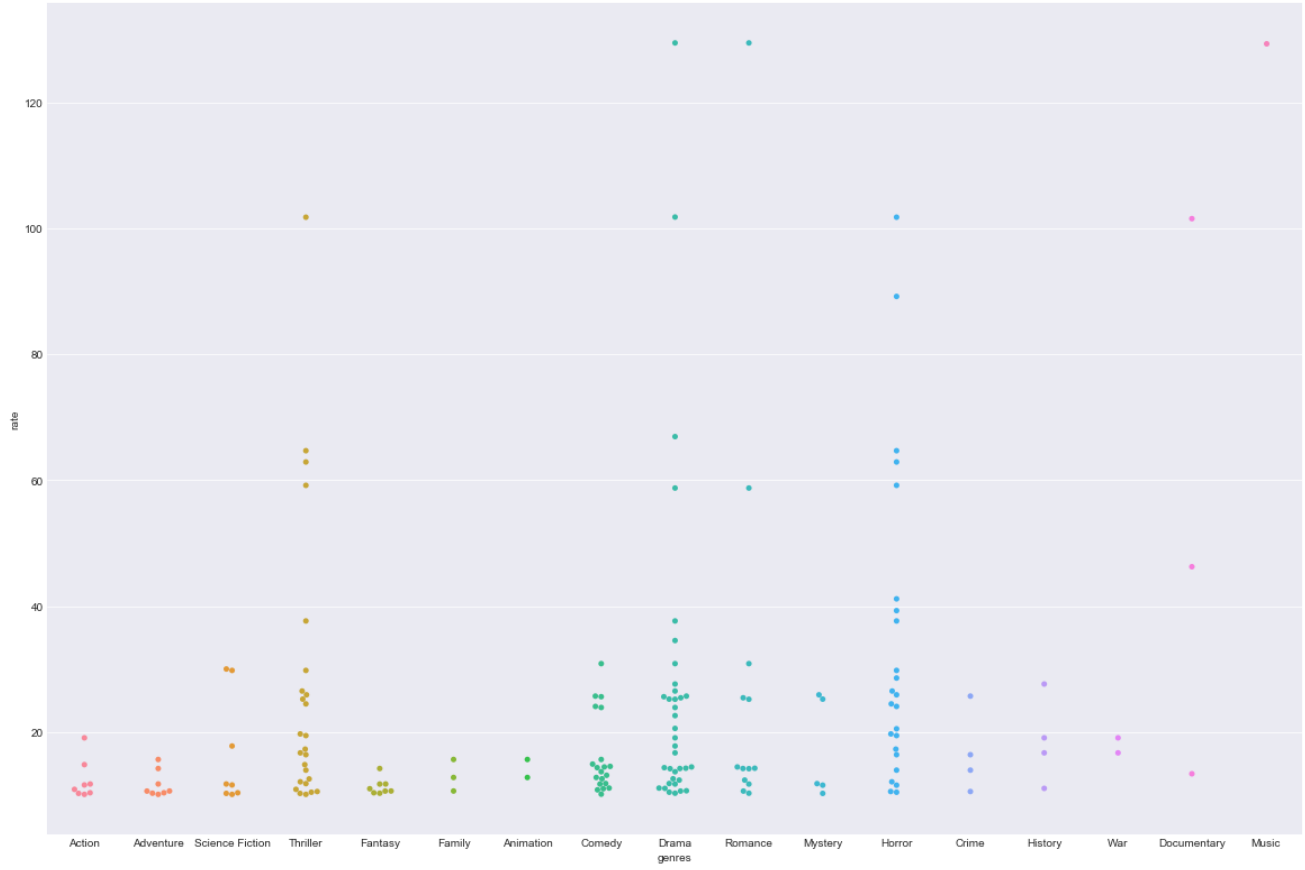
	id	imdb_id	popularity	original_title	cast	director	tagline
6821	18925	tt0805526	0.206454	Facing the Giants	James Blackwell|Alex Kendrick|Shannen Fields|C...	Alex Kendrick	Never give up. Never back down. Never lose faith.
7437	5723	tt0907657	1.254438	Once	Glen Hansard|MarkÃ©ta IrglovÃ¡|Hugh Walsh|Gera...	John Carney	How often do you find the right person?
7437	5723	tt0907657	1.254438	Once	Glen Hansard|MarkÃ©ta IrglovÃ¡|Hugh Walsh|Gera...	John Carney	How often do you find the right person?
7437	5723	tt0907657	1.254438	Once	Glen Hansard|MarkÃ©ta IrglovÃ¡|Hugh Walsh|Gera...	John Carney	How often do you find the right person?
242	299245	tt2309260	0.532205	The Gallows	Cassidy Gifford|Ryan Shoos|Pfeifer Brown|Reese...	Travis Cluff|Chris Lofing	Every School Has Its Spirit
242	299245	tt2309260	0.532205	The Gallows	Cassidy Gifford|Ryan Shoos|Pfeifer Brown|Reese...	Travis Cluff|Chris Lofing	Every School Has Its Spirit
3581	59296	tt1436559	0.520430	Love, Wedding, Marriage	Mandy Moore|Kellan Lutz|Jessica Szohr|Autumn F...	Dermot Mulroney	Here comes the ride.
3581	59296	tt1436559	0.520430	Love, Wedding, Marriage	Mandy Moore|Kellan Lutz|Jessica Szohr|Autumn F...	Dermot Mulroney	Here comes the ride.
7447	23827	tt1179904	1.120442	Paranormal Activity	Katie Featherston|Micah Sloat|Mark Fredrichs|A...	Oren Peli	What Happens When You Sleep?
7447	23827	tt1179904	1.120442	Paranormal Activity	Katie Featherston|Micah Sloat|Mark Fredrichs|A...	Oren Peli	What Happens When You Sleep?
3608	50217	tt0893412	0.463510	From Prada to Nada	Camilla Belle|Alexa PenaVega|April Bowlby|Wilm...	Angel Gracia	A riches to rags story.
3608	50217	tt0893412	0.463510	From Prada to Nada	Camilla Belle|Alexa PenaVega|April Bowlby|Wilm...	Angel Gracia	A riches to rags story.
3608	50217	tt0893412	0.463510	From Prada to Nada	Camilla Belle|Alexa PenaVega|April Bowlby|Wilm...	Angel Gracia	A riches to rags story.

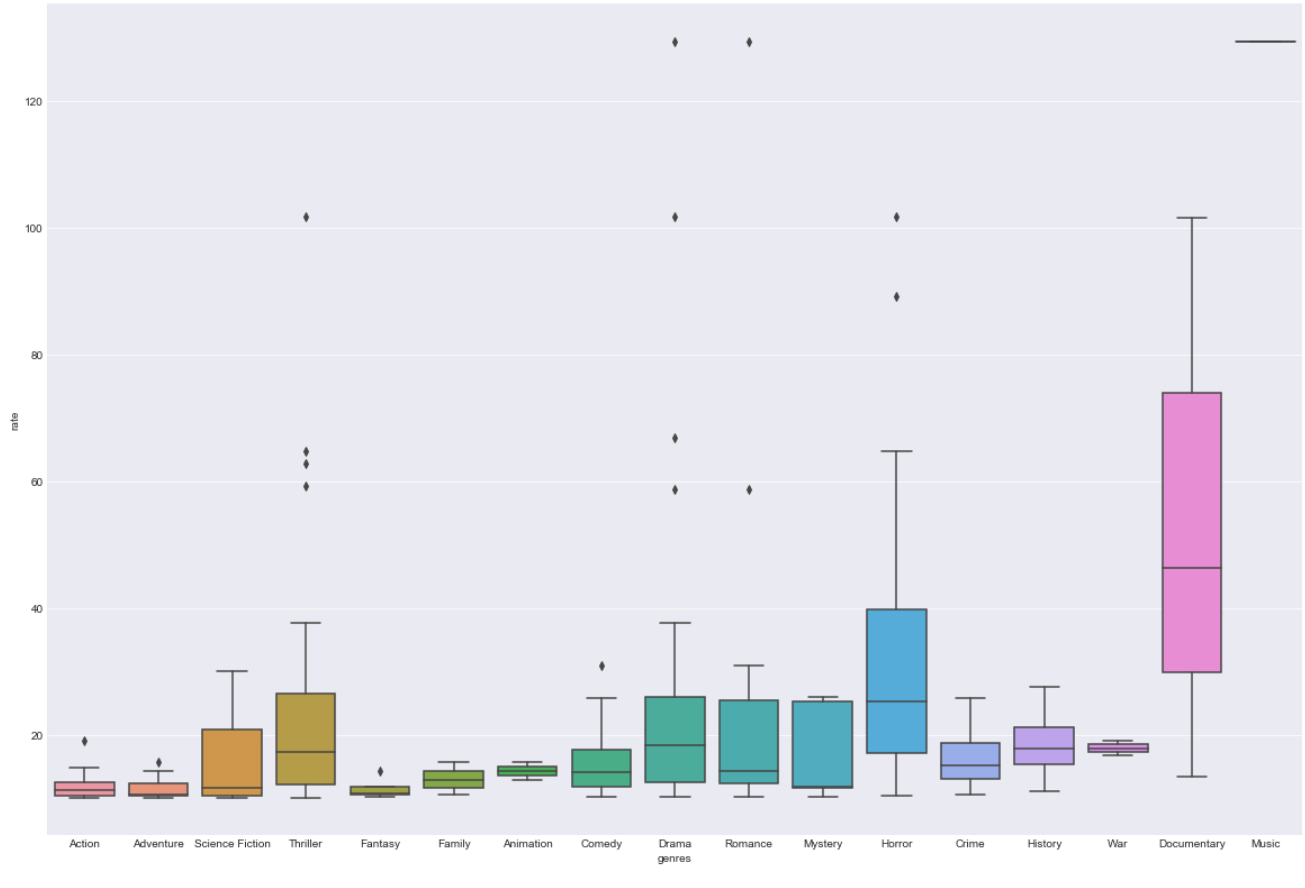
4438 rows × 17 columns

In [13]:
```
#筛选出票房收入是预算十倍以上的电影
df_3 = df_3[df_3['rate'] > 10]
#绘制散点图和箱线图
ax = plt.subplots(figsize=(20,30))
plt.subplot(211)
sns.swarmplot(df_3['genres'], df_3['rate'], data=df_3)
plt.subplot(212)
sns.boxplot(df_3['genres'], df_3['rate'], data=df_3);
```



In [14]:
```
#去掉过大的异常值，使得图像结果易于观察
index_3 = df_3[df_3['rate'] > 400].index.tolist()
df_3 = df_3.drop(index_3)
#绘制散点图和箱线图
ax = plt.subplots(figsize=(20,30))
plt.subplot(211)
sns.swarmplot(df_3['genres'], df_3['rate'], data=df_3)
plt.subplot(212)
sns.boxplot(df_3['genres'], df_3['rate'], data=df_3);
```





结合散点图和箱线图，恐怖惊悚类型的电影以及喜剧、浪漫和戏剧类型的电影最有以较低的成本获取较高票房收入的潜力，并且第一步去掉的四个异常值的电影类型分别是恐怖惊悚类型和喜剧浪漫类型的电影。

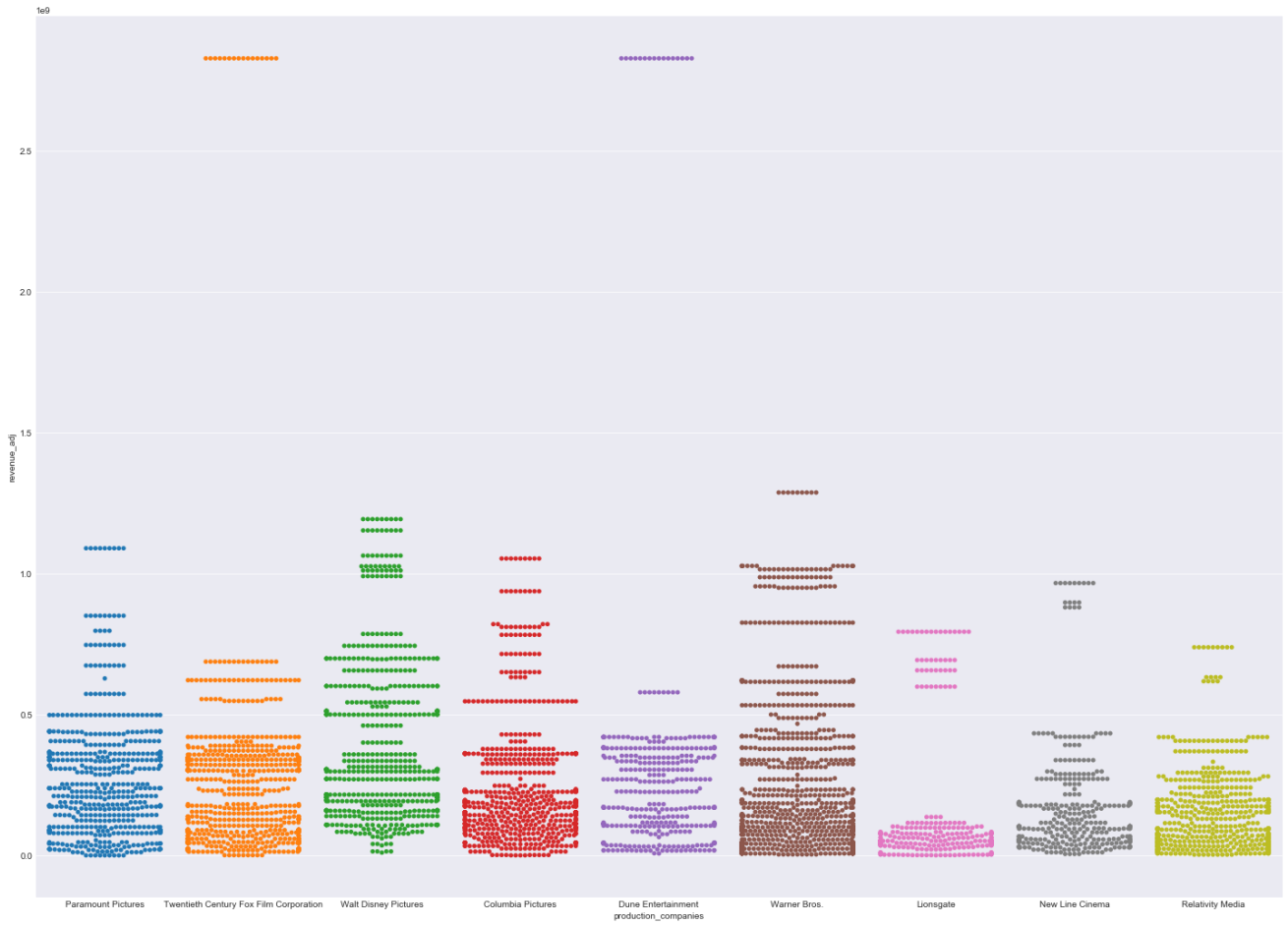
4. 哪些公司的电影更能制作出受欢迎的电影?

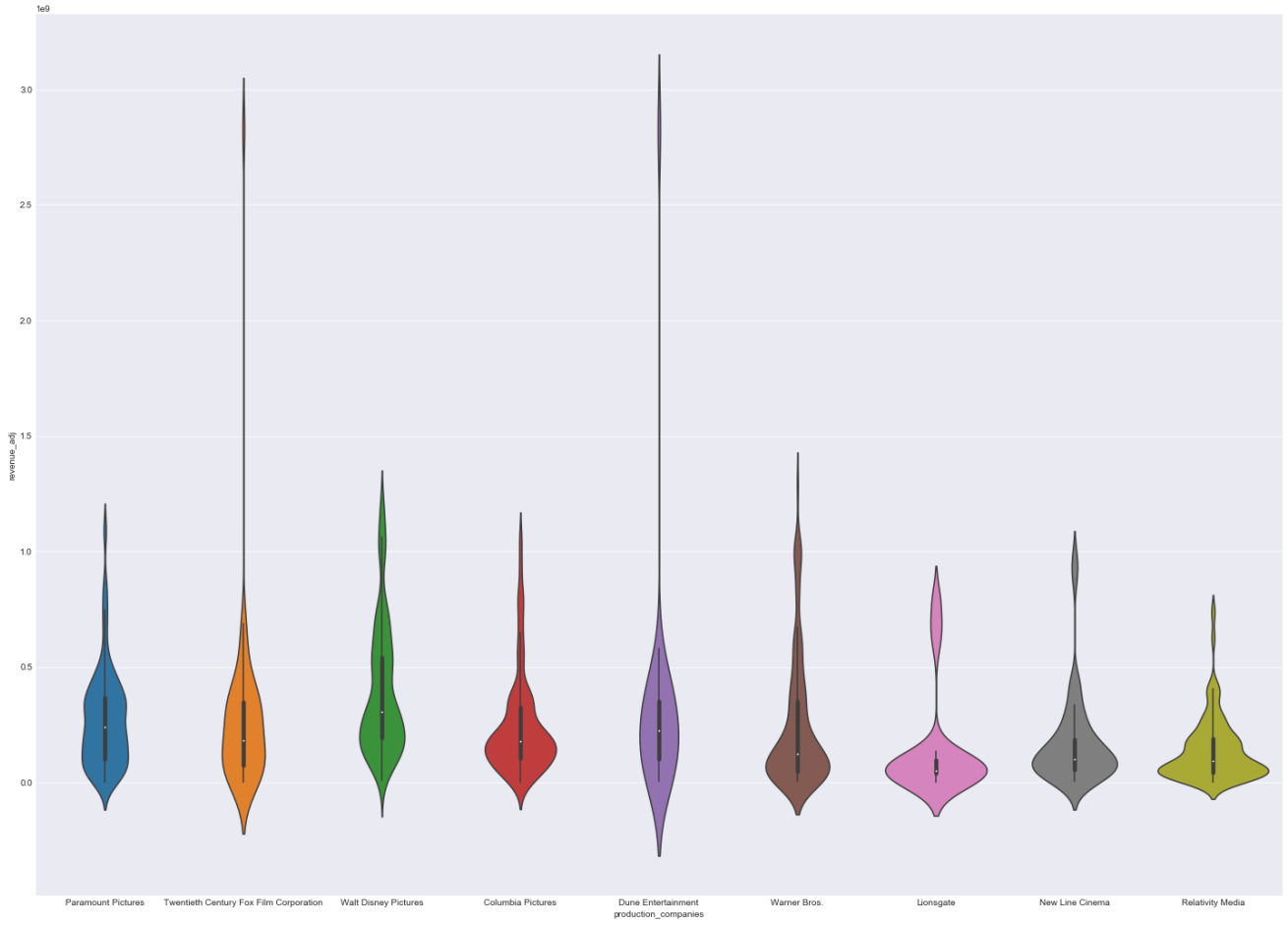
In [15]:
```
df_4 = df_10years
#将production_companies列的类型关键词拆分，并合成新的一列
df_4 = df_4.drop('production_companies', axis=1).join(df_4['production_companies'].str.split(
'|', expand=True).stack().reset_index(level=1,drop=True).rename('production_companies'))
#观察主要的电影制作公司有哪些
df_4['production_companies'].value_counts()
```

Out[15]:
```
Warner Bros.                                    1108
Universal Pictures                               968
Twentieth Century Fox Film Corporation           738
Columbia Pictures                                719
Relativity Media                                 686
Walt Disney Pictures                             625
Paramount Pictures                               604
Summit Entertainment                             410
Dune Entertainment                               395
Legendary Pictures                               373
New Line Cinema                                  365
Regency Enterprises                              332
Village Roadshow Pictures                        330
Lionsgate                                        313
DreamWorks Animation                             268
Canal+                                           257
Fox 2000 Pictures                                248
Ingenious Film Partners                          229
StudioCanal                                      213
Millennium Films                                 212
Screen Gems                                      209
Metro-Goldwyn-Mayer (MGM)                        205
Vertigo Entertainment                            191
Touchstone Pictures                              188
Original Film                                    180
DreamWorks SKG                                   179
Lakeshore Entertainment                          179
Twentieth Century Fox Animation                  178
TriStar Pictures                                 172
Di Bonaventura Pictures                          168
                                                 ...
Cott Productions                                   1
Fantastic Films                                    1
Julijette                                          1
Movie Package Company (MPC)                        1
Backup Media                                       1
Hearst Entertainment Productions                   1
TelevisiÃ³ de Catalunya TV3                        1
Underground                                        1
Noruz Films                                        1
Ce Qui Me Meut Motion Pictures                     1
Prime Focus Ltd.                                   1
Collision Entertainment                            1
Gulfstream Pictures                                1
Fewlas Entertainment                               1
Sony Pictures Worldwide Acquisitions (SPWA)        1
Solar Filmworks                                    1
Battle Mountain Films                              1
TMC Films                                          1
Toma 78                                            1
Little Stranger                                    1
Bernard Gayle Productions                          1
Banter                                             1
Blue Rider Pictures                                1
Zucker/Netter Productions                          1
Michaels-Goldwyn                                   1
Dickhouse Productions                              1
Jack and Henry Productions Inc.                    1
In A World                                         1
MPI Media Group                                    1
Enderby Entertainment                              1
Name: production_companies, Length: 1954, dtype: int64
```

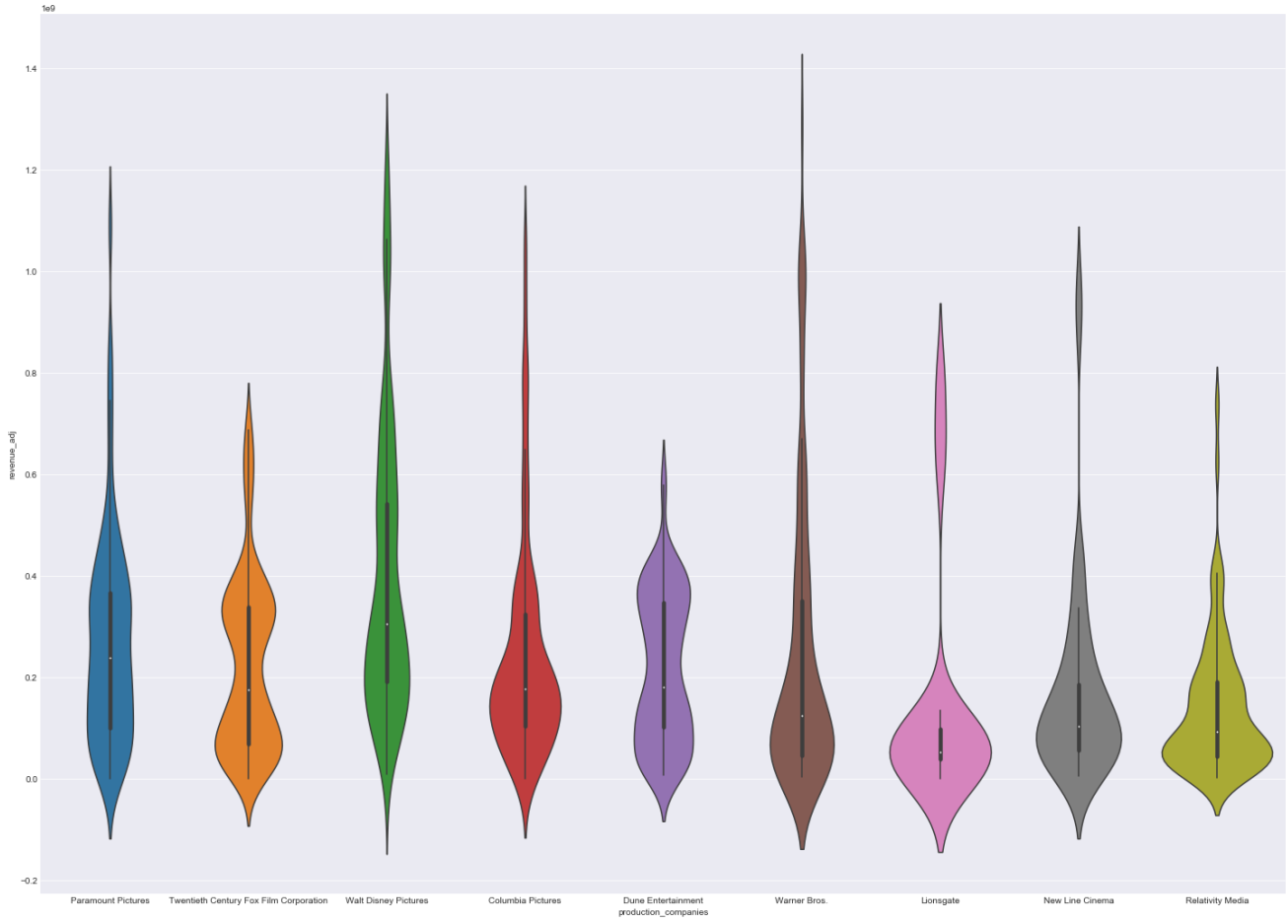
In [16]:
```
#挑选出十个主要的电影制作公司
df_10companies = df_4[(df_4['production_companies'] == 'Universal Pictures ')|
                      (df_4['production_companies'] == 'Warner Bros.')|
                      (df_4['production_companies'] == 'Relativity Media')|
                      (df_4['production_companies'] == 'Columbia Pictures')|
                      (df_4['production_companies'] == 'Paramount Pictures')|
                      (df_4['production_companies'] == 'Twentieth Century Fox Film Corporatio
n')|
                      (df_4['production_companies'] == 'New Line Cinema')|
                      (df_4['production_companies'] == 'Walt Disney Pictures')|
                      (df_4['production_companies'] == 'Dune Entertainment')|
                      (df_4['production_companies'] == 'Lionsgate')]
```

In [17]:
```
#绘制散点图和提琴图
ax = plt.subplots(figsize=(25,40))
plt.subplot(211)
sns.swarmplot(df_10companies['production_companies'], df_10companies['revenue_adj'],data=df_10
companies)
plt.subplot(212)
sns.violinplot(df_10companies['production_companies'], df_10companies['revenue_adj'],data=df_1
0companies);
```


1e9
2.5
2.0
1.5
1.0
0.5
0.0
revenue_adj
Paramount Pictures
Twentieth Century Fox Film Corporation
Walt Disney Pictures
Columbia Pictures
Dune Entertainment
Warner Bros.
Lionsgate
New Line Cinema
Relativity Media
production_companies


1e9
3.0
2.5
2.0
1.5
1.0
0.5
0.0
revenue_adj
Paramount Pictures
Twentieth Century Fox Film Corporation
Walt Disney Pictures
Columbia Pictures
Dune Entertainment
Warner Bros.
Lionsgate
New Line Cinema
Relativity Media
production_companies

In [18]:
```
#去掉最大的异常值
df_10companies = df_10companies[df_10companies['revenue_adj'] < 1.5e9]
#绘制散点图和提琴图
ax = plt.subplots(figsize=(25,40))
plt.subplot(211)
sns.swarmplot(df_10companies['production_companies'], df_10companies['revenue_adj'],data=df_10
companies)
plt.subplot(212)
sns.violinplot(df_10companies['production_companies'], df_10companies['revenue_adj'],data=df_1
0companies);
```



结合散点图和提琴图，相比之下，华纳兄弟、派拉蒙电影公司、华特迪士尼以及沙丘娱乐这四家电影公司更有能力制作出大众喜爱的电影作品。

结论

1. 近十年来，最受欢迎的电影类型是动作冒险类，其次是喜剧、戏剧以及恐怖类和奇幻类型的电影。并且动作冒险类越来越受欢迎。
2. 电影的票房收入和预算较强显的正相关关系，但是由于未进行统计检验以及本研究的局限性，无法得出预算越高票房就越高的结论；
   而票房和电影平均评分有一定的正相关关系但是相关关系不强，也就是说评分高的电影不一定会有较高的票房收入。
3. 各大电影类型中恐怖惊悚类型的电影以及喜剧、浪漫和戏剧类型的电影最有以较低的成本获取较高票房收入的潜力。
4. 近年来，众多电影制作公司中，华纳兄弟、派拉蒙电影公司、华特迪士尼以及沙丘娱乐这四家电影公司更有能力制作出大众喜爱的电影作品。