Student Name: Hassan Abdullah

Student ID: 2131117616

Exploratory Data Analysis

Perform exploratory analysis to find some initial insights on the following data sets:

- movies.json (https://raw.githubusercontent.com/vega/vega-datasets/gh-pages/data/movies.json)

Remember, that you are approaching the data with no specific question, only to get some general insights on it, so you can be able to ask the right questions in future analysis.

Be sure to perform the following steps:

1. Identify the variables in the data set and prepare a table describing what each variable represents. See table markdown (https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet#tables) to see how to write create markdown tables in your report. Description should include:
    - Variable definition
    - Data type
    - Missing data report
    - Report on the distribution of the data
    - level of analysis
2. Include table or list of all transformed variables/aggregations that were used in the study and include:
    - Variable description
    - Steps in transformation
    - Distribution if applicable
    - level of analysis
3. Start exploring relationships and groups to identify insights. Under every graph, write the main insights derived from the graph, and then compile a list of insights at the top of the report
4. Prepare a list of possible questions that come to mind after discovering these insights, and explain whether the question can be answered with the current data, or will require more data?

Note: Include responses to these 4 items in the top 4 cells of the report using mardown. the analysis should be at the bottom of the report in a section labeled **Analysis**

Report:

Variables	Description	Data type	Notes

Variables	Description	Data type	Notes
Creative_Type	if the movie have fixtion idea an whats is it	object, categorical	
Director	directer name	object, categorical	
Distributor	Distributor name	object, categorical	
IMDB_Rating	movie rating on IMDB site	float64, categorical	
IMDB_Votes	movie votes on IMDB site	float64, continuous	
MPAA_Rating	can be describes as parents Guidelines (pictures rating)	object, categorical	
Major_Genre	genre of the movie	object, categorical	
Production_Budget	budget of production	object, continuous	
Release_Date	movie release date to the cinema	datetime, categorical	i changed type from float64 to date time
Rotten_Tomatoes_Rating	movie rating on Rotten Tomatoes site	float64, categorical	
Source	Source of the movie idea	object,categorical	
Running_time_min	movie duration in minutes	float64, continuous	
Title	title of the movie	object, categorical	
US_DVD_Sales	dvd sales in the united states	float64, continuous	
US_Gross	total sales in the united states	float64, continuous	
Worldwide_Gross	total sales around the world	float64, continuous	

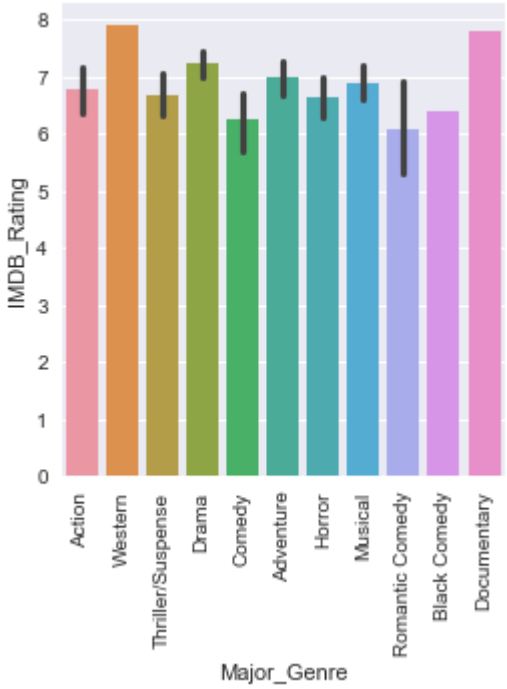
- there was some missing data in he data frame that needed to be filtered to clean_movies_df
- Ridley Scott is the director with the most movies in the data frame.(4 movies)
- there are 174 creative_type in the data frame 8 of them are unique and the top is Contemporary Fiction with freq of 75.
    - count 174
    - unique 8
    - top Contemporary Fiction
    - freq 75
- there are 174 director 147 of them are unique, top director is Ridley Scott with 4 movies.
    - count 174
    - unique 147
    - top Ridley Scott

    - freq 4
- there are 174 release date 128 of them are unique, the top date is 15-dec-06 with freq of 3.
    - count 174
    - unique 128
    - top 15-Dec-06
    - freq 3
- there are 174 rotten tomatoes rating with mean of 55.7, min of 2, max of 98.
    - count 174.000000
    - mean 55.724138
    - std 27.973641
    - min 2.000000
    - 25% 29.000000
    - 50% 61.000000
    - 75% 79.000000
    - max 98.000000
- there are 174 running time with mean of 114.344, min 78, and max of 187.
    - count 174.000000
    - mean 114.344828
    - std 20.116612
    - min 78.000000
    - 25% 99.250000
    - 50% 111.000000
    - 75% 126.750000
    - max 187.000000
- there are 174 source, 10 of them are unique, the top source is Original Screenplay with freq of 75.
    - count 174
    - unique 10
    - top Original Screenplay
    - freq 75
- there are 174 title, all of them are unique.
    - count 174
    - unique 174
    - top Elizabethtown
    - freq 1
- there are 174 value of US_DVD_Sales with mean(53,923,880), min(618,454), and max(320,830,900)
    - count 1.740000e+02
    - mean 5.392388e+07
    - std 6.322215e+07
    - min 6.184540e+05
    - 25% 1.492860e+07
    - 50% 2.813295e+07
    - 75% 6.764799e+07

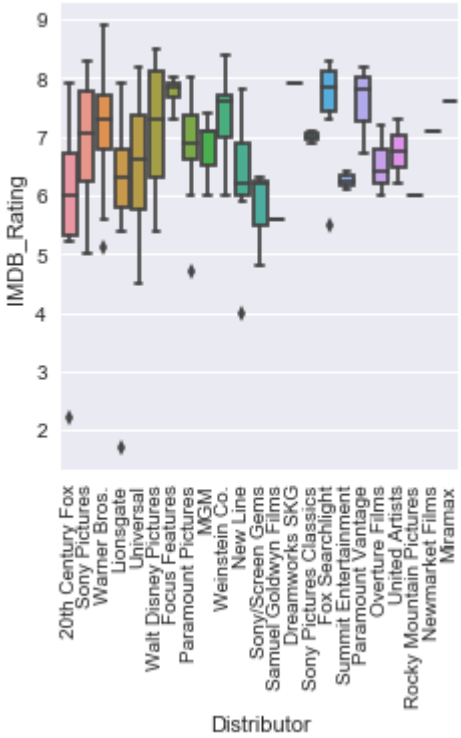
    - max 3.208309e+08
- there are 174 value of US_Gross with mean(93,863,240), min(2,223,293), and max(533,345,400)
    - count 1.740000e+02
    - mean 9.386324e+07
    - std 9.707097e+07
    - min 2.223293e+06
    - 25% 2.575447e+07
    - 50% 5.483773e+07
    - 75% 1.276575e+08
    - max 5.333454e+08
- there are 174 value of Worldwide_Gross with mean(207,487,300), min(6,521,829), max(1,065,660,000)
    - count 1.740000e+02
    - mean 2.074873e+08
    - std 2.354148e+08
    - min 6.521829e+06
    - 25% 4.895248e+07
    - 50% 1.187267e+08
    - 75% 2.651107e+08
    - max 1.065660e+09
- there are 174 distributor 23 of them are unique with freq of 24.
    - count 174
    - unique 23
    - top Universal
    - freq 24
- there are 174 value of IMDB_Rating with mean(6.78), min(1.7), max(8.9)
    - count 174.000000
    - mean 6.788506
    - std 1.093959
    - min 1.700000
    - 25% 6.100000
    - 50% 7.000000
    - 75% 7.600000
    - max 8.900000
- there are 174 value of IMDB_Votes with mean(62767), min(149), max(465000)
    - count 174.000000
    - mean 62767.545977
    - std 60418.301378
    - min 149.000000
    - 25% 20571.000000
    - 50% 44370.500000
    - 75% 84531.000000
    - max 465000.000000

- there are 174 value of MPAA_Rating 4 of them re unique the top value is PG-13
    - count 174
    - unique 4
    - top PG-13
    - freq 76
- there are 174 Major Genre 11 of them are unique, the top freq is Drama (46)
    - count 174
    - unique 11
    - top Drama
    - freq 46
- there are 174 value of Production_Budget with mean(66,802,960), min(15,000), max(300,000,000)
    - count 1.740000e+02
    - mean 6.680296e+07
    - std 6.008128e+07
    - min 1.500000e+04
    - 25% 2.225000e+07
    - 50% 4.900000e+07
    - 75% 8.500000e+07
    - max 3.000000e+08

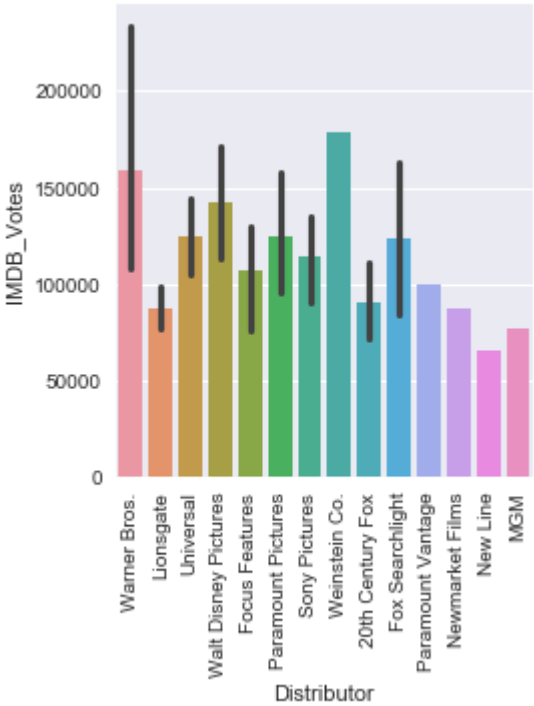
- Highest rating for western and decumentary.



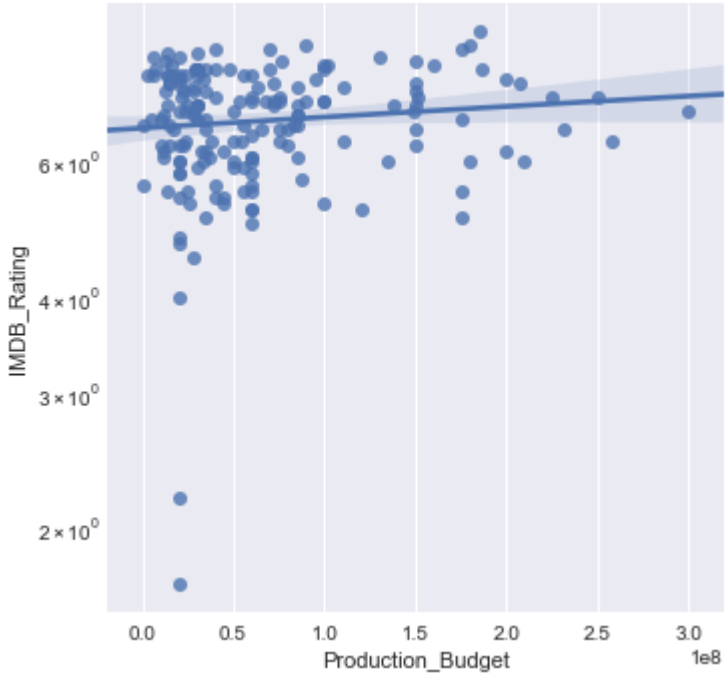
- Highest mean of rating is for Focus Fuetures.
- lowest Rating mean is for 20th century fox.



- warner Bros. have the highest votes number.
- new line have the least votes number.



- High production budget doesn't mean its successful movie.(there are no clear relation between budgent and rate)



Analysis:

In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
movies_df = pd.read_json("https://raw.githubusercontent.com/vega/vega-datasets/gh
```

In [2]:
```
sns.set(rc={'figure.figsize':(11.7,8.27)})
```

In [23]:
```
movies_df
```

Out[23]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating
0	None	None	Gramercy	6.1	1071.0	R
1	None	None	Strand	6.9	207.0	R
2	None	None	Lionsgate	6.8	865.0	None
3	None	None	Fine Line	NaN	NaN	None
4	Contemporary Fiction	None	Trimark	3.4	165.0	R
5	None	None	MGM	NaN	NaN	None
6	None	Christopher Nolan	Zeitgeist	7.7	15133.0	R

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

Out[8]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Genre	Pro
0	None	None	Gramercy	6.1	1071.0	R	None	
1	None	None	Strand	6.9	207.0	R	Drama	
2	None	None	Lionsgate	6.8	865.0	None	Comedy	
3	None	None	Fine Line	NaN	NaN	None	Comedy	
4	Contemporary Fiction	None	Trimark	3.4	165.0	R	Drama	

In [39]:
```
movies_df.isnull().any(axis=1)
```

Out[39]:
```
0        True
1        True
2        True
3        True
4        True
5        True
6        True
7        True
8        True
9        True
10       True
11       True
12       True
13       True
14       True
15       True
16       True
17       True
18       True
19       True
20       True
21       True
22       True
23       True
24       True
25       True
26       True
27       True
28       True
29       True
        ...
3171     True
3172     True
3173     True
3174     True
3175     True
3176     True
3177     True
3178     True
3179     True
3180     True
3181    False
3182     True
3183    False
3184     True
3185     True
3186     True
3187     True
3188     True
3189     True
3190     True
3191     True
3192     True
3193     True
3194     True
```

```
3195    False
3196    False
3197     True
3198     True
3199     True
3200     True
Length: 3201, dtype: bool
```

In [3]:
```
clean_movies_df = movies_df[~(movies_df.isnull().any(axis=1)) | (movies_df.duplic
```

In [45]:
```
clean_movies_df.columns
```

Out[45]:
```
Index(['Creative_Type', 'Director', 'Distributor', 'IMDB_Rating', 'IMDB_Votes',
       'MPAA_Rating', 'Major_Genre', 'Production_Budget', 'Release_Date',
       'Rotten_Tomatoes_Rating', 'Running_Time_min', 'Source', 'Title',
       'US_DVD_Sales', 'US_Gross', 'Worldwide_Gross'],
      dtype='object')
```

In [47]:
```
clean_movies_df.Creative_Type.describe()
```

Out[47]:
```
count                     174
unique                      8
top       Contemporary Fiction
freq                       75
Name: Creative_Type, dtype: object
```

In [48]:
```
clean_movies_df.Director.describe()
```

Out[48]:
```
count              174
unique             147
top       Ridley Scott
freq                 4
Name: Director, dtype: object
```

In [49]:
```
clean_movies_df.Distributor.describe()
```

Out[49]:
```
count           174
unique           23
top       Universal
freq             24
Name: Distributor, dtype: object
```

In [50]:
```
clean_movies_df.IMDB_Rating.describe()
```

Out[50]:
```
count    174.000000
mean       6.788506
std        1.093959
min        1.700000
25%        6.100000
50%        7.000000
75%        7.600000
max        8.900000
Name: IMDB_Rating, dtype: float64
```

In [51]:
```
clean_movies_df.IMDB_Votes.describe()
```

Out[51]:
```
count       174.000000
mean      62767.545977
std       60418.301378
min         149.000000
25%       20571.000000
50%       44370.500000
75%       84531.000000
max      465000.000000
Name: IMDB_Votes, dtype: float64
```

In [52]:
```
clean_movies_df.MPAA_Rating.describe()
```

Out[52]:
```
count       174
unique        4
top       PG-13
freq         76
Name: MPAA_Rating, dtype: object
```

In [53]:
```
clean_movies_df.Major_Genre.describe()
```

Out[53]:
```
count       174
unique       11
top       Drama
freq         46
Name: Major_Genre, dtype: object
```

In [4]:
```
clean_movies_df.Production_Budget.describe()
```

Out[4]:
```
count    1.740000e+02
mean     6.680296e+07
std      6.008128e+07
min      1.500000e+04
25%      2.225000e+07
50%      4.900000e+07
75%      8.500000e+07
max      3.000000e+08
Name: Production_Budget, dtype: float64
```

In [5]:
```
clean_movies_df.Release_Date.describe()
```

Out[5]:
```
count           174
unique          128
top       15-Dec-06
freq              3
Name: Release_Date, dtype: object
```

In [6]:
```
clean_movies_df.Rotten_Tomatoes_Rating.describe()
```

Out[6]:
```
count    174.000000
mean      55.724138
std       27.973641
min        2.000000
25%       29.000000
50%       61.000000
75%       79.000000
max       98.000000
Name: Rotten_Tomatoes_Rating, dtype: float64
```

In [7]:
```
clean_movies_df.Running_Time_min.describe()
```

Out[7]:
```
count    174.000000
mean     114.344828
std       20.116612
min       78.000000
25%       99.250000
50%      111.000000
75%      126.750000
max      187.000000
Name: Running_Time_min, dtype: float64
```

In [8]:
```
clean_movies_df.Source.describe()
```

Out[8]:
```
count                     174
unique                     10
top       Original Screenplay
freq                       75
Name: Source, dtype: object
```

In [9]:
```
clean_movies_df.Title.describe()
```

Out[9]:
```
count               174
unique              174
top       Elizabethtown
freq                  1
Name: Title, dtype: object
```

In [10]:
```
clean_movies_df.US_DVD_Sales.describe()
```

Out[10]:
```
count    1.740000e+02
mean     5.392388e+07
std      6.322215e+07
min      6.184540e+05
25%      1.492860e+07
50%      2.813295e+07
75%      6.764799e+07
max      3.208309e+08
Name: US_DVD_Sales, dtype: float64
```

In [11]:
```
clean_movies_df.US_Gross.describe()
```

Out[11]:
```
count    1.740000e+02
mean     9.386324e+07
std      9.707097e+07
min      2.223293e+06
25%      2.575447e+07
50%      5.483773e+07
75%      1.276575e+08
max      5.333454e+08
Name: US_Gross, dtype: float64
```

In [12]:
```
clean_movies_df.Worldwide_Gross.describe()
```

Out[12]:
```
count    1.740000e+02
mean     2.074873e+08
std      2.354148e+08
min      6.521829e+06
25%      4.895248e+07
50%      1.187267e+08
75%      2.651107e+08
max      1.065660e+09
Name: Worldwide_Gross, dtype: float64
```

In [13]:
```
clean_movies_df["Release_Date"]= pd.to_datetime(clean_movies_df.Release_Date)
```

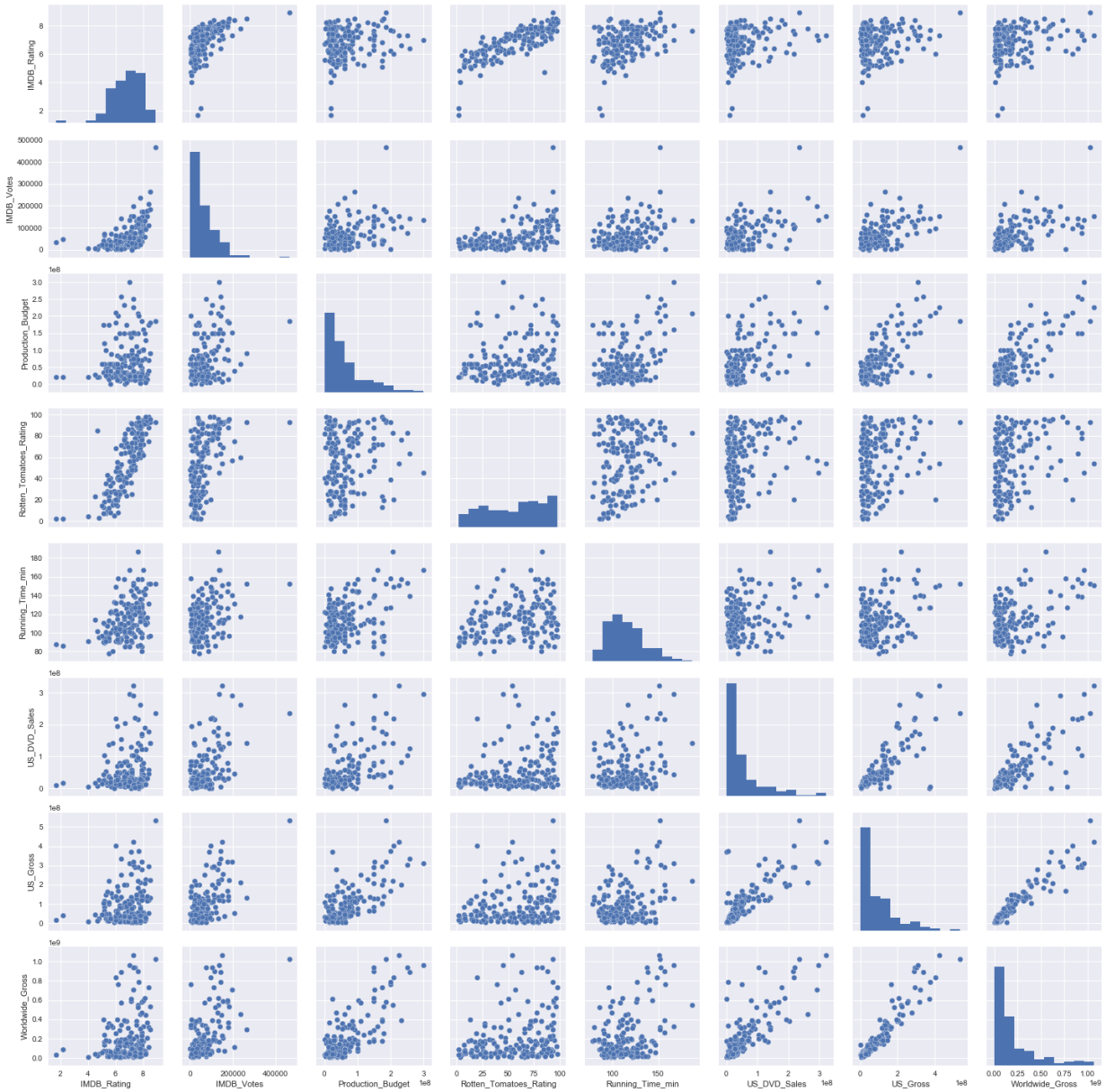
```
C:\Users\almousawi\Anaconda3\lib\site-packages\ipykernel_launcher.py:1: Setting
WithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-docs/stab
le/indexing.html#indexing-view-versus-copy (http://pandas.pydata.org/pandas-doc
s/stable/indexing.html#indexing-view-versus-copy)
  """Entry point for launching an IPython kernel.
```

In [17]:
```
clean_movies_df.dtypes
```

Out[17]:
```
Creative_Type                     object
Director                          object
Distributor                       object
IMDB_Rating                      float64
IMDB_Votes                       float64
MPAA_Rating                       object
Major_Genre                       object
Production_Budget                float64
Release_Date              datetime64[ns]
Rotten_Tomatoes_Rating           float64
Running_Time_min                 float64
Source                            object
Title                             object
US_DVD_Sales                     float64
US_Gross                         float64
Worldwide_Gross                  float64
dtype: object
```

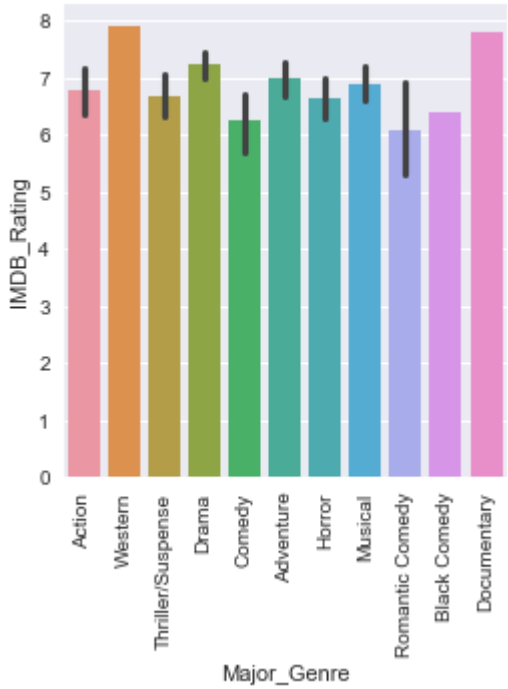
In [61]:
```
sns.pairplot(clean_movies_df)
```

Out[61]: <seaborn.axisgrid.PairGrid at 0x873aa41fd0>



In [66]:
```
sns.factorplot(x="Major_Genre" , y="IMDB_Rating", data=clean_movies_df, kind="bar
plt.xticks(rotation= 90)
```

Out[66]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10]),
 <a list of 11 Text xticklabel objects>)
```



In [85]:
```
mean_rating = clean_movies_df.IMDB_Rating.mean()
```

In [95]:
```
clean_movies_df[clean_movies_df.IMDB_Rating > 8]
```

Out[95]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Ge
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	Ac
1355	Contemporary Fiction	Paul Greengrass	Universal	8.2	146025.0	PG-13	Ac
1391	Historical Fiction	Quentin Tarantino	Weinstein Co.	8.4	178742.0	R	Ac
1448	Science Fiction	Alfonso Cuaron	Universal	8.1	158125.0	R	Thriller/Suspe
1591	Science Fiction	Neill Blomkamp	Sony Pictures	8.3	151742.0	R	Thriller/Suspe
1616	Contemporary Fiction	Martin Scorsese	Warner Bros.	8.5	264148.0	R	Dra
2163	Historical Fiction	Clint Eastwood	Warner Bros.	8.1	56872.0	R	Dra
2446	Contemporary Fiction	Sean Penn	Paramount Vantage	8.2	99464.0	R	Dra
2566	Historical Fiction	Christopher Nolan	Walt Disney Pictures	8.4	207322.0	PG-13	Thriller/Suspe
2596	Fantasy	Brad Bird	Walt Disney Pictures	8.1	131929.0	G	Com
2774	Contemporary Fiction	Danny Boyle	Fox Searchlight	8.3	176325.0	R	Dra
3056	Kids Fiction	Pete Docter	Walt Disney Pictures	8.4	110491.0	PG	Adven
3095	Kids Fiction	Andrew Stanton	Walt Disney Pictures	8.5	182257.0	G	Com
3158	Contemporary Fiction	Darren Aronofsky	Fox Searchlight	8.2	93301.0	R	Dra

In [106]:
```
clean_movies_df[clean_movies_df.Director == "Christopher Nolan"]
```

Out[106]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Ge
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	Ac
2566	Historical Fiction	Christopher Nolan	Walt Disney Pictures	8.4	207322.0	PG-13	Thriller/Suspe

In [111]:
```
clean_movies_df.Director.value_counts()
```

Out[111]:
```
Ridley Scott            4
Sam Raimi               3
Bryan Singer            2
Adam Shankman           2
Kevin MacDonald         2
Gore Verbinski          2
Anne Fletcher           2
James Mangold           2
Jason Friedberg         2
Darren Lynn Bousman     2
Christopher Nolan       2
Mel Gibson              2
Renny Harlin            2
Sam Fell                2
Todd Phillips           2
Michael Bay             2
Sam Mendes              2
Danny Boyle             2
Darren Aronofsky        2
Jon Avnet               2
D.J. Caruso             2
David Yates             2
Catherine Hardwicke     2
Marc Forster            2
Matthew Vaughn          1
Nick Park               1
Michael O. Sajbel       1
Les Mayfield            1
Martin Campbell         1
Greg Mottola            1
                       ..
Antoine Fuqua           1
Joel Schumacher         1
Joe Carnahan            1
Tom Tykwer              1
Andy Fickman            1
Roland Emmerich         1
David R. Ellis          1
Martin Scorsese         1
Terry Gilliam           1
Louis Leterrier         1
Alex Proyas             1
Clint Eastwood          1
Tim Hill                1
Rob Minkoff             1
Oren Peli               1
Russell Mulcahy         1
Andrew Stanton          1
Phil Lord               1
Gary Winick             1
Jon Favreau             1
Quentin Tarantino       1
Michael Winterbottom    1
Brad Bird               1
Alex Kendrick           1
```

```
Frank Darabont          1
Jonathan Demme          1
Curtis Hanson           1
Edgar Wright            1
David Fincher           1
Peyton Reed             1
Name: Director, Length: 147, dtype: int64
```

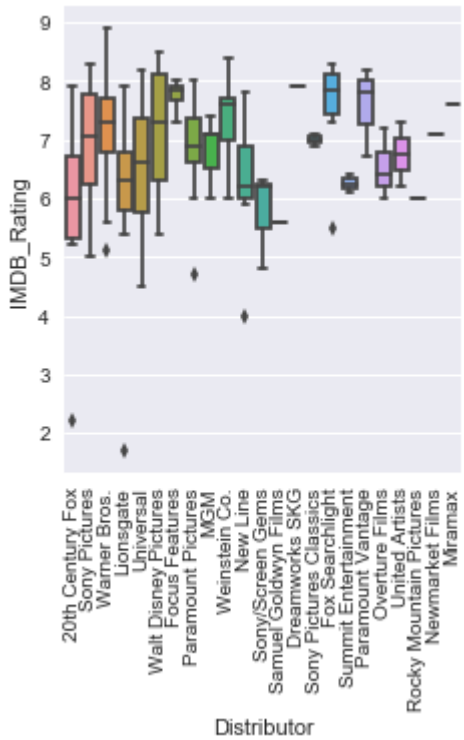
In [118]:
```
clean_movies_df.sort_values(by=["IMDB_Rating"], ascending=False).head(10)
```

Out[118]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Ge
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	Ac
1616	Contemporary Fiction	Martin Scorsese	Warner Bros.	8.5	264148.0	R	Dra
3095	Kids Fiction	Andrew Stanton	Walt Disney Pictures	8.5	182257.0	G	Com
3056	Kids Fiction	Pete Docter	Walt Disney Pictures	8.4	110491.0	PG	Adven
1391	Historical Fiction	Quentin Tarantino	Weinstein Co.	8.4	178742.0	R	Ac
2566	Historical Fiction	Christopher Nolan	Walt Disney Pictures	8.4	207322.0	PG-13	Thriller/Suspe
2774	Contemporary Fiction	Danny Boyle	Fox Searchlight	8.3	176325.0	R	Dra
1591	Science Fiction	Neill Blomkamp	Sony Pictures	8.3	151742.0	R	Thriller/Suspe
1355	Contemporary Fiction	Paul Greengrass	Universal	8.2	146025.0	PG-13	Ac
2446	Contemporary Fiction	Sean Penn	Paramount Vantage	8.2	99464.0	R	Dra

In [285]:
```
sns.factorplot(data=clean_movies_df, x="Distributor", y="IMDB_Rating", kind="box"
plt.xticks(rotation= 90)
```

Out[285]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13, 14, 15, 16,
        17, 18, 19, 20, 21, 22]), <a list of 23 Text xticklabel objects>)
```



In [265]:
```
clean_movies_df.Distributor.value_counts()
```

Out[265]:
```
Universal                 24
Warner Bros.              23
Sony Pictures             22
Paramount Pictures        20
20th Century Fox          14
Walt Disney Pictures      13
New Line                   9
Lionsgate                  9
Fox Searchlight            6
MGM                        5
Weinstein Co.              5
Focus Features             4
Paramount Vantage          3
Overture Films             3
Sony/Screen Gems           3
United Artists             2
Sony Pictures Classics     2
Summit Entertainment       2
Samuel Goldwyn Films       1
Dreamworks SKG             1
Newmarket Films            1
Miramax                    1
Rocky Mountain Pictures    1
Name: Distributor, dtype: int64
```

In [142]:
```
clean_movies_df.Distributor.describe()
```

Out[142]:
```
count           174
unique           23
top       Universal
freq             24
Name: Distributor, dtype: object
```

In [144]:
```
clean_movies_df[clean_movies_df.Distributor == "Universal"]
```

Out[144]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_
1147	Fantasy	Tom Shadyac	Universal	5.5	43164.0	PG	C
1165	Dramatization	Ridley Scott	Universal	7.9	114060.0	R	
1355	Contemporary Fiction	Paul Greengrass	Universal	8.2	146025.0	PG-13	
1448	Science Fiction	Alfonso Cuaron	Universal	8.1	158125.0	R	Thriller/Sus
1502	Contemporary Fiction	Peter Billingsley	Universal	5.5	18332.0	PG-13	C
1541	Historical Fiction	Brian De Palma	Universal	5.6	35210.0	R	Thriller/Sus
1618	Science Fiction	Paul Anderson	Universal	6.6	40611.0	R	
1619	Fantasy	Sam Raimi	Universal	7.1	51343.0	PG-13	
1795	Dramatization	Ron Howard	Universal	7.9	36366.0	R	
1808	Contemporary Fiction	Garry Marshall	Universal	5.8	10902.0	R	
1992	Super Hero	Louis Leterrier	Universal	7.1	82419.0	PG-13	Adv
2036	Contemporary Fiction	Dennis Dugan	Universal	6.1	46347.0	PG-13	C
2060	Dramatization	Sam Mendes	Universal	7.2	60650.0	R	
2123	Fantasy	Peter Jackson	Universal	7.6	132720.0	PG-13	Adv
2124	Contemporary Fiction	Judd Apatow	Universal	7.5	111192.0	R	Ro C
2126	Contemporary Fiction	Peter Berg	Universal	7.1	47200.0	R	
2147	Fantasy	Brad Silberling	Universal	5.3	16830.0	PG-13	C
2371	Fantasy	Rob Cohen	Universal	5.1	41570.0	PG-13	Adv
2380	Contemporary Fiction	Michael Mann	Universal	6.0	51921.0	R	

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_
2778	Contemporary Fiction	Joe Carnahan	Universal	6.6	57313.0	R	C
2871	Contemporary Fiction	Kevin MacDonald	Universal	7.3	34067.0	PG-13	Thriller/Sus
2930	Kids Fiction	Sam Fell	Universal	6.1	7460.0	G	Adv
3126	Contemporary Fiction	Malcolm D. Lee	Universal	4.5	5700.0	PG-13	C
3146	Science Fiction	Timur Bekmambetov	Universal	6.4	1089.0	R	

In [145]:
```
clean_movies_df[clean_movies_df.Distributor == "Dreamworks SKG"]
```

Out[145]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Genre
1872	Kids Fiction	Nick Park	Dreamworks SKG	7.9	38158.0	G	Adventure

In [154]:
```
above_average_rating = clean_movies_df[clean_movies_df.IMDB_Rating > mean_rating]
```

In [199]:
```
above_average_rating.sort_values(by=["IMDB_Rating"], ascending=False)
```

Out[199]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	
3095	Kids Fiction	Andrew Stanton	Walt Disney Pictures	8.5	182257.0	G	C
1616	Contemporary Fiction	Martin Scorsese	Warner Bros.	8.5	264148.0	R	
2566	Historical Fiction	Christopher Nolan	Walt Disney Pictures	8.4	207322.0	PG-13	Thriller/Su
3056	Kids Fiction	Pete Docter	Walt Disney Pictures	8.4	110491.0	PG	Adv
1391	Historical Fiction	Quentin Tarantino	Weinstein Co.	8.4	178742.0	R	

In [186]:
```
above_average_rating.Director.value_counts()
```

Out[186]:
```
Ridley Scott           4
Sam Mendes             2
Sam Raimi              2
Kevin MacDonald        2
David Yates            2
Christopher Nolan      2
Danny Boyle            2
Marc Forster           2
Mel Gibson             2
Darren Aronofsky       2
Gore Verbinski         2
James Mangold          2
Ruben Fleischer        1
Steven Soderbergh      1
Stephen Daldry         1
Frank Darabont         1
Jonathan Demme         1
Paul Haggis            1
Bryan Singer           1
Richard LaGravenese    1
Pete Docter            1
Edward Zwick           1
Kevin Smith            1
Todd Field             1
Edgar Wright           1
David Fincher          1
Rob Marshall           1
Neill Blomkamp         1
Larry Charles          1
Darren Lynn Bousman    1
                      ..
F. Gary Gray           1
Antoine Fuqua          1
Peter Jackson          1
Tom Tykwer             1
Emilio Estevez         1
D.J. Caruso            1
Martin Scorsese        1
Terry Gilliam          1
Louis Leterrier        1
Clint Eastwood         1
James Gray             1
Kevin Lima             1
Greg Mottola           1
Denzel Washington      1
Zack Snyder            1
Mike Newell            1
Ben Stiller            1
Judd Apatow            1
Stephen Frears         1
Gavin Hood             1
Michael Bay            1
Nick Cassavetes        1
Alfonso Cuaron         1
Len Wiseman            1
```

```
Ron Howard             1
Sam Fell               1
Jason Reitman          1
Spike Jonze            1
Martin Campbell        1
Peyton Reed            1
Name: Director, Length: 83, dtype: int64
```

In [187]:
```
above_average_rating.Director.describe()
```

Out[187]:
```
count               97
unique              83
top       Ridley Scott
freq                 4
Name: Director, dtype: object
```

In [203]:
```
radiley_scott = clean_movies_df[clean_movies_df.Director == "Ridley Scott"]
```

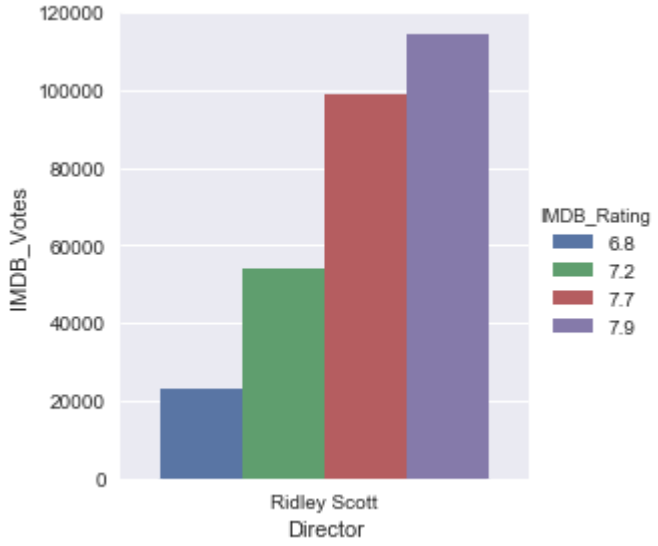
In [201]:
```
clean_movies_df[clean_movies_df.Director == "Ridley Scott"]
```

Out[201]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Genr
1127	Contemporary Fiction	Ridley Scott	20th Century Fox	6.8	23149.0	PG-13	Drama
1165	Dramatization	Ridley Scott	Universal	7.9	114060.0	R	Drama
1278	Contemporary Fiction	Ridley Scott	Warner Bros.	7.2	53921.0	R	Thriller/Suspens
1306	Dramatization	Ridley Scott	Sony Pictures	7.7	98653.0	R	Actio

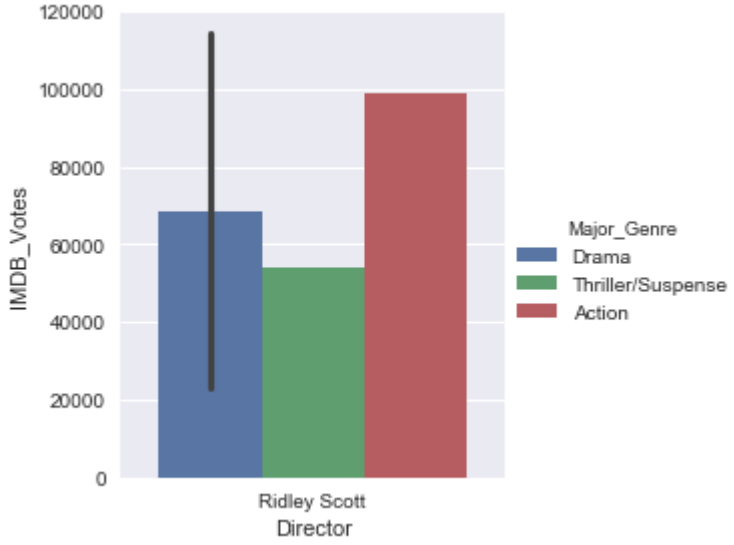
In [210]:
```
sns.factorplot(data=radiley_scott, x="Director", y="IMDB_Votes", hue="IMDB_Rating
```

Out[210]:
```
<seaborn.axisgrid.FacetGrid at 0x8750e04710>
```



In [215]:
```
sns.factorplot(data=radiley_scott, x="Director", y="IMDB_Votes", hue="Major_Genre
```

Out[215]:
```
<seaborn.axisgrid.FacetGrid at 0x87513eef60>
```



In [222]:
```
mean_votes = clean_movies_df.IMDB_Votes.mean()
```

In [223]:
```
above_mean_votes = clean_movies_df[clean_movies_df.IMDB_Votes > mean_votes]
```

In [228]:
```
above_mean_votes.sort_values(["IMDB_Votes"], ascending=[False])
```

Out[228]:

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	
1616	Contemporary Fiction	Martin Scorsese	Warner Bros.	8.5	264148.0	R	D
1090	Historical Fiction	Zack Snyder	Warner Bros.	7.8	235508.0	R	
2566	Historical Fiction	Christopher Nolan	Walt Disney Pictures	8.4	207322.0	PG-13	Thriller/Sus
2940	Science Fiction	Michael Bay	Paramount Pictures	7.3	197131.0	PG-13	
3095	Kids Fiction	Andrew Stanton	Walt Disney	8.5	182257.0	G	Co

In [225]:
```
above_mean_votes.Distributor.describe()
```

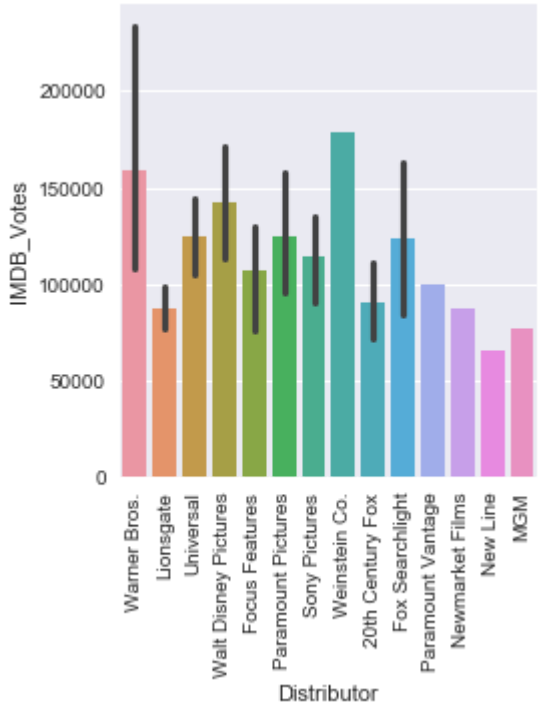
Out[225]:
```
count               62
unique              14
top       Warner Bros.
freq                12
Name: Distributor, dtype: object
```

In [226]:
```
above_mean_votes.Distributor.value_counts()
```

Out[226]:
```
Warner Bros.            12
Sony Pictures           10
Walt Disney Pictures     7
Paramount Pictures       7
Universal                6
20th Century Fox         6
Fox Searchlight          4
Focus Features           3
Lionsgate                2
Paramount Vantage        1
Newmarket Films          1
Weinstein Co.            1
MGM                      1
New Line                 1
Name: Distributor, dtype: int64
```

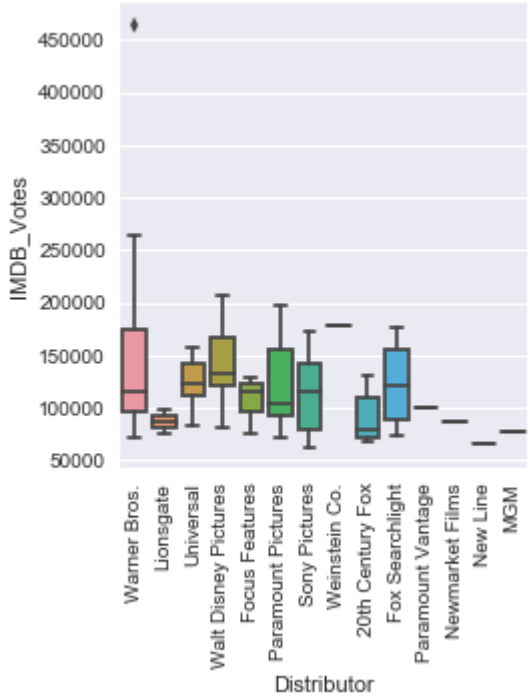
In [237]:
```
sns.factorplot(data=above_mean_votes, x="Distributor", y="IMDB_Votes", kind="bar"
plt.xticks(rotation=90)
```

Out[237]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
 <a list of 14 Text xticklabel objects>)
```



In [289]:
```
sns.factorplot(data=above_mean_votes, x="Distributor", y="IMDB_Votes", kind="box"
plt.xticks(rotation=90)
```

Out[289]:
```
(array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11, 12, 13]),
 <a list of 14 Text xticklabel objects>)
```



In [256]:
```
warner_bros = clean_movies_df[clean_movies_df.Distributor == "Warner Bros."]
```

In [257]:
```
warner_bros.sort_values(by="IMDB_Votes", ascending=False)
```

Out[257]:

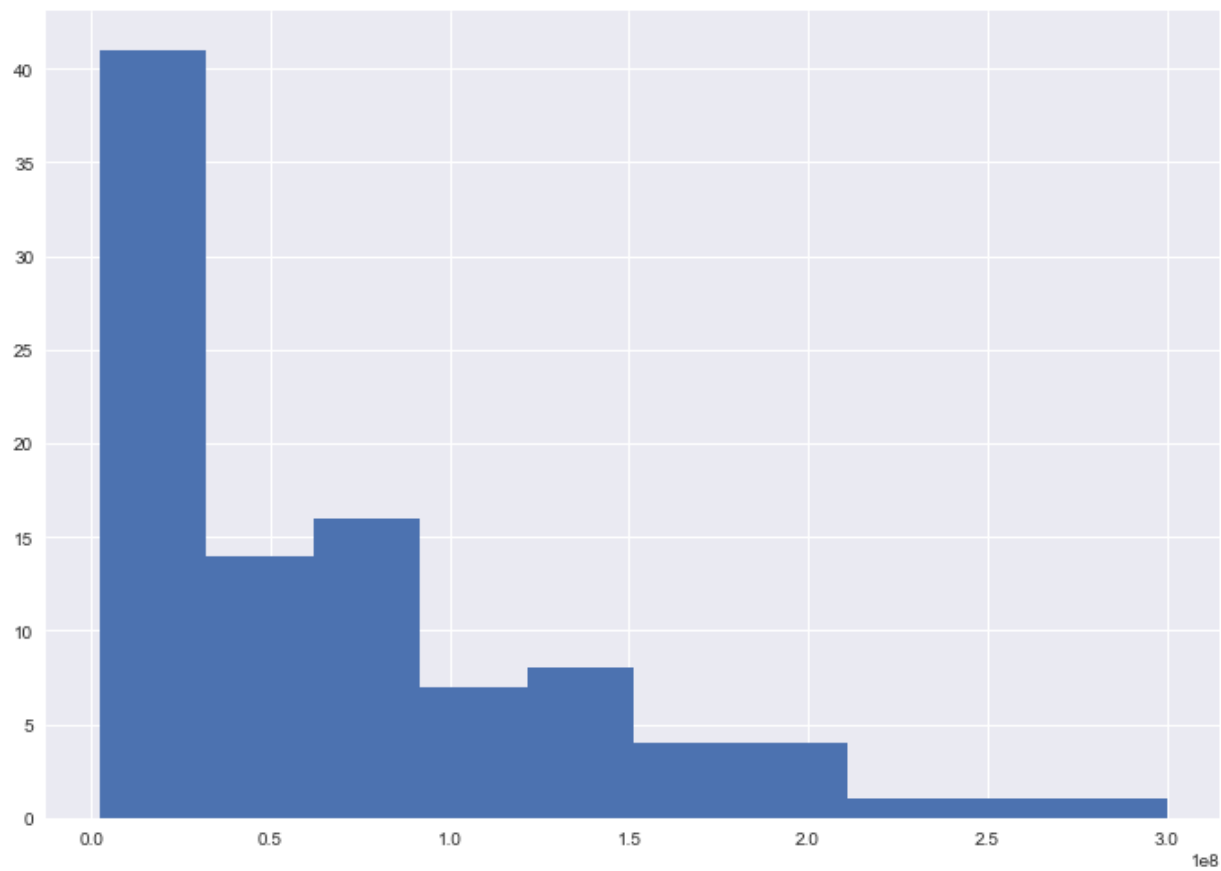
	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Ge
1266	Super Hero	Christopher Nolan	Warner Bros.	8.9	465000.0	PG-13	Ac
1616	Contemporary Fiction	Martin Scorsese	Warner Bros.	8.5	264148.0	R	Dra
1090	Historical Fiction	Zack Snyder	Warner Bros.	7.8	235508.0	R	Ac
2160	Science Fiction	Francis Lawrence	Warner Bros.	7.1	153631.0	PG-13	Ho
1946	Contemporary Fiction	Todd Phillips	Warner Bros.	7.9	127634.0	R	Com
1279	Historical Fiction	Edward Zwick	Warner Bros.	8.0	118925.0	R	Ac
1972	Fantasy	Mike Newell	Warner Bros.	7.6	111946.0	PG-13	Adven
1973	Fantasy	David Yates	Warner Bros.	7.4	104074.0	PG-13	Adven
2828	Super Hero	Bryan Singer	Warner Bros.	6.6	102751.0	PG-13	Adven
2454	Contemporary Fiction	Steven Soderbergh	Warner Bros.	6.9	76884.0	PG-13	Adven
1974	Fantasy	David Yates	Warner Bros.	7.3	73720.0	PG	Adven
1773	Science Fiction	Darren Aronofsky	Warner Bros.	7.4	72562.0	PG-13	Dra
3183	Contemporary Fiction	Peyton Reed	Warner Bros.	7.0	62150.0	PG-13	Com
2163	Historical Fiction	Clint Eastwood	Warner Bros.	8.1	56872.0	R	Dra
1278	Contemporary Fiction	Ridley Scott	Warner Bros.	7.2	53921.0	R	Thriller/Suspe
1976	Kids Fiction	George Miller	Warner Bros.	6.7	42369.0	PG	Adven

	Creative_Type	Director	Distributor	IMDB_Rating	IMDB_Votes	MPAA_Rating	Major_Ge
3131	Fantasy	Spike Jonze	Warner Bros.	7.2	30669.0	PG	Adven
3068	Dramatization	Paul Haggis	Warner Bros.	7.4	27529.0	R	Dra
2610	Fantasy	Stephen Hopkins	Warner Bros.	5.6	19881.0	R	Ho
2218	Contemporary Fiction	Ken Kwapis	Warner Bros.	5.1	15422.0	PG	Roma Com
2034	Fantasy	Iain Softley	Warner Bros.	6.1	14157.0	PG	Adven
2391	Contemporary Fiction	Nick Cassavetes	Warner Bros.	7.4	13839.0	PG-13	Dra
2224	Historical Fiction	Curtis Hanson	Warner Bros.	5.9	9870.0	PG-13	Dra

In [268]:
```
production_budget_df = above_average_rating.sort_values(by=["Production_Budget"],
```

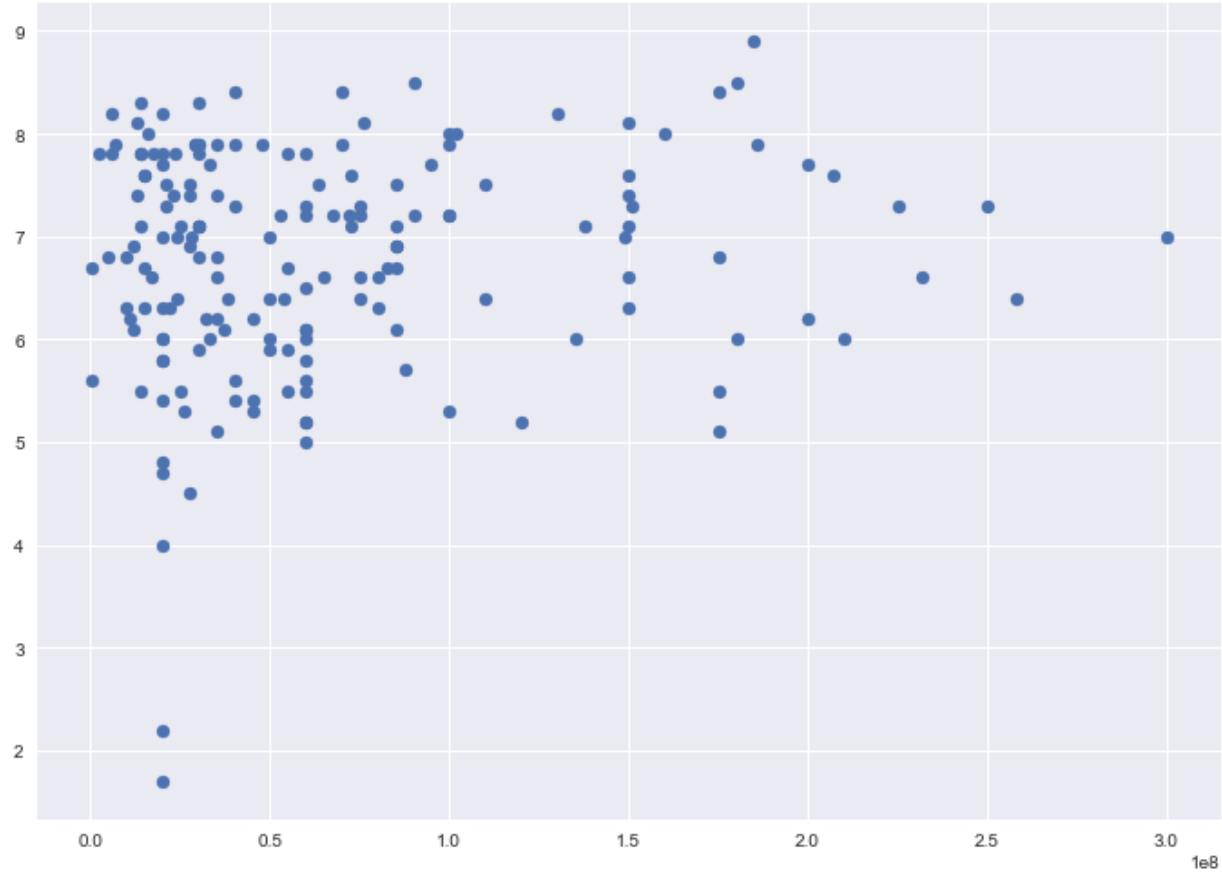
In [290]:
```
production_budget_df.Production_Budget.hist()
```

Out[290]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x875800aef0>
```



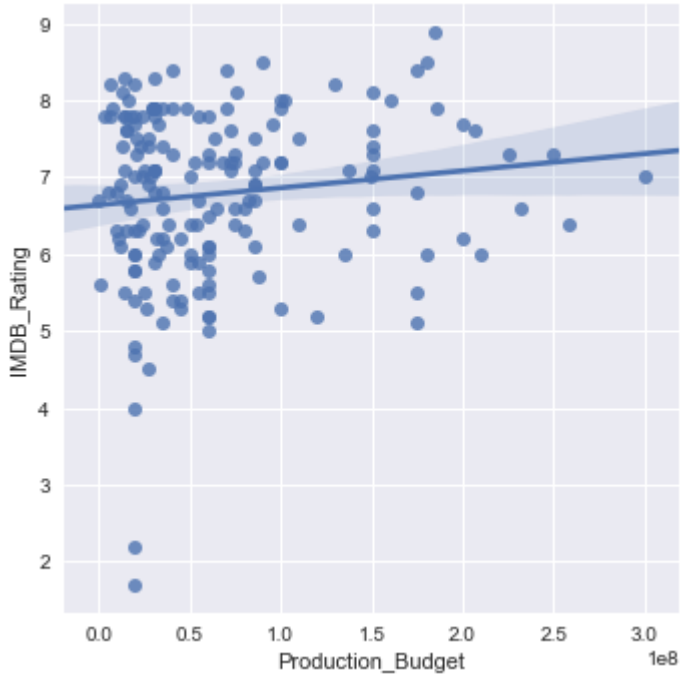
In [286]:
```
plt.scatter(data=clean_movies_df, x="Production_Budget", y="IMDB_Rating")
```

Out[286]: <matplotlib.collections.PathCollection at 0x87584dc2b0>



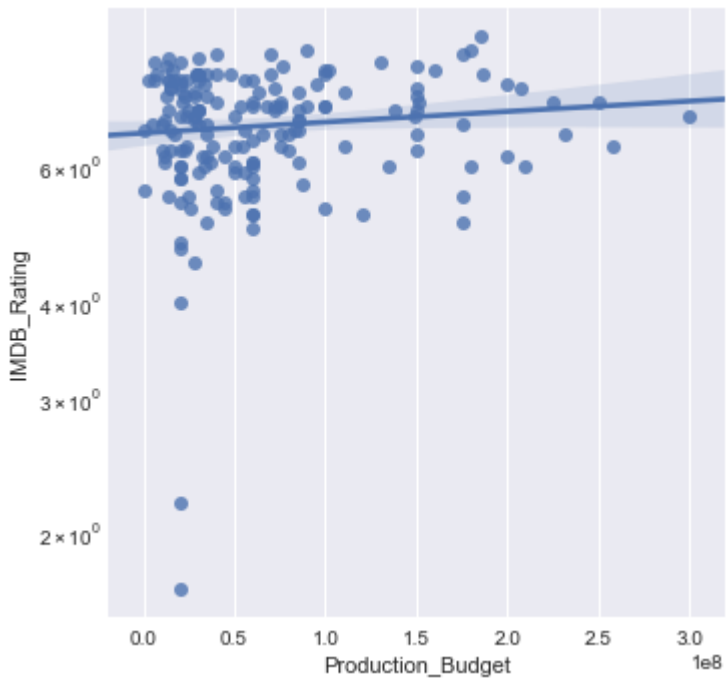
In [291]:
```
sns.lmplot(data=clean_movies_df, x="Production_Budget", y="IMDB_Rating")
```

Out[291]: <seaborn.axisgrid.FacetGrid at 0x8754d5b898>



In [303]:
```
h = sns.lmplot(data=clean_movies_df, x="Production_Budget", y="IMDB_Rating")
h.set(yscale="log")
```

Out[303]: <seaborn.axisgrid.FacetGrid at 0x875dacc7f0>



In [ ]:

In [ ]: