Correlation-on-Movie-Dataset

January 19, 2024

0.0.1 About Dataset:

0.0.2 Context:

The dataset is focused on movie revenue and factors that affects the movie over the last decades (6820 movies in the dataset (220 movies per year, 1986-2016)).

0.0.3 Contents:

The dataset includes Budget, Company, Country, Director, Genre, Gross, Name, Rating, Runtime, Score(IMDb rating), Votes, Star Cast(lead actors), Writer, Year Relased.

0.0.4 Acknowledgements

This data was scraped from IMDb.

0.0.5 Problem:

Analyse the movie factors and the correlation between them to find the cause of slow-down of movie industry.

Import Libraries and dataset

[1]:
```
# Import Libraries
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab
```

[2]:
```
# Import Data
df=pd.read_csv(r"C:\Users\amitm\Desktop\New folder\Project\Python␣
 ↪Projects\movies.csv")
```

[3]:
```
df.head()
```

[3]:

	name	rating	genre	year	\
0	The Shining	R	Drama	1980	
1	The Blue Lagoon	R	Adventure	1980	
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	

```
3                                    Airplane!     PG     Comedy  1980
4                                   Caddyshack      R     Comedy  1980

                         released  score      votes         director  \
0  June 13, 1980 (United States)    8.4   927000.0  Stanley Kubrick
1   July 2, 1980 (United States)    5.8    65000.0   Randal Kleiser
2  June 20, 1980 (United States)    8.7  1200000.0   Irvin Kershner
3   July 2, 1980 (United States)    7.7   221000.0     Jim Abrahams
4  July 25, 1980 (United States)    7.3   108000.0     Harold Ramis

                    writer            star         country      budget  \
0            Stephen King  Jack Nicholson  United Kingdom  19000000.0
1  Henry De Vere Stacpoole  Brooke Shields   United States   4500000.0
2          Leigh Brackett     Mark Hamill   United States  18000000.0
3            Jim Abrahams     Robert Hays   United States   3500000.0
4       Brian Doyle-Murray     Chevy Chase   United States   6000000.0

         gross             company  runtime
0   46998772.0        Warner Bros.    146.0
1   58853106.0   Columbia Pictures    104.0
2  538375067.0           Lucasfilm    124.0
3   83453539.0  Paramount Pictures     88.0
4   39846344.0      Orion Pictures     98.0
```

[4]:
```
# Data Description
df.describe()
```

[4]:
```
              year        score         votes        budget         gross  \
count  7668.000000  7665.000000  7.665000e+03  5.497000e+03  7.479000e+03
mean   2000.405451     6.390411  8.810850e+04  3.558988e+07  7.850054e+07
std      11.153508     0.968842  1.633238e+05  4.145730e+07  1.657251e+08
min    1980.000000     1.900000  7.000000e+00  3.000000e+03  3.090000e+02
25%    1991.000000     5.800000  9.100000e+03  1.000000e+07  4.532056e+06
50%    2000.000000     6.500000  3.300000e+04  2.050000e+07  2.020576e+07
75%    2010.000000     7.100000  9.300000e+04  4.500000e+07  7.601669e+07
max    2020.000000     9.300000  2.400000e+06  3.560000e+08  2.847246e+09

           runtime
count  7664.000000
mean    107.261613
std      18.581247
min      55.000000
25%      95.000000
50%     104.000000
75%     116.000000
max     366.000000
```

[5]:
```
print (df.dtypes)
```

```
name         object
rating       object
genre        object
year          int64
released     object
score       float64
votes       float64
director     object
writer       object
star         object
country      object
budget      float64
gross       float64
company      object
runtime     float64
dtype: object
```

Data Cleaning and Manipulation

[6]:
```
# Data Cleaning and Manipulation

# Creating a copy of our dataset.
df1 = df.copy()
df1
```

[6]:

	name	rating	genre	year	released	score	votes	director
0	The Shining	R	Drama	1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner
3	Airplane!	PG	Comedy	1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams
4	Caddyshack	R	Comedy	1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis
…	…	…	…	…	…	…	…	…
7663	More to Life	NaN	Drama	2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks
7664	Dream Round	NaN	Comedy	2020				
7665	Saving Mbango	NaN	Drama	2020				
7666	It's Just Us	NaN	Drama	2020				
7667	Tee em el	NaN	Horror	2020				

```
7664  February 7, 2020 (United States)    4.7       36.0     Dusty Dukatz
7665         April 27, 2020 (Cameroon)    5.7       29.0     Nkanya Nkwai
7666   October 1, 2020 (United States)    NaN        NaN    James Randall
7667   August 19, 2020 (United States)    5.7        7.0     Pereko Mosia

                      writer             star         country      budget  \
0               Stephen King    Jack Nicholson  United Kingdom  19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States   4500000.0
2              Leigh Brackett       Mark Hamill   United States  18000000.0
3                Jim Abrahams       Robert Hays   United States   3500000.0
4          Brian Doyle-Murray       Chevy Chase   United States   6000000.0
…                         …                 …               …          …
7663             Joseph Ebanks      Shannon Bond   United States      7000.0
7664               Lisa Huston  Michael Saquella   United States         NaN
7665              Lynno Lovert      Onyama Laura   United States     58750.0
7666             James Randall     Christina Roz   United States     15000.0
7667              Pereko Mosia  Siyabonga Mabaso    South Africa         NaN

           gross                    company  runtime
0      46998772.0                Warner Bros.    146.0
1      58853106.0           Columbia Pictures    104.0
2     538375067.0                    Lucasfilm    124.0
3      83453539.0          Paramount Pictures     88.0
4      39846344.0              Orion Pictures     98.0
…             …                           …       …
7663          NaN                         NaN     90.0
7664          NaN  Cactus Blue Entertainment     90.0
7665          NaN            Embi Productions      NaN
7666          NaN                         NaN    120.0
7667          NaN                 PK 65 Films    102.0

[7668 rows x 15 columns]
```

[7]:
```
# Manipulation of Coloumn
y = df1['released'].str.replace(")","").str.split("(",expand=True).
  ↪rename(columns={0:'released_date',1:'released_place'})
y
```

[7]:
```
          released_date released_place
0        June 13, 1980   United States
1         July 2, 1980   United States
2        June 20, 1980   United States
3         July 2, 1980   United States
4        July 25, 1980   United States
…                  …              …
7663  October 23, 2020   United States
7664  February 7, 2020   United States
```

```
7665    April 27, 2020        Cameroon
7666   October 1, 2020   United States
7667   August 19, 2020   United States

[7668 rows x 2 columns]
```

[8]:
```
df1.insert(4,'released_date',y['released_date'])
df1
```

[8]:

	name	rating	genre	year	released_date	released	score	votes	director	writer	star
0	The Shining	R	Drama	1980	June 13, 1980	June 13, 1980 (United States)	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson
1	The Blue Lagoon	R	Adventure	1980	July 2, 1980	July 2, 1980 (United States)	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	June 20, 1980	June 20, 1980 (United States)	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill
3	Airplane!	PG	Comedy	1980	July 2, 1980	July 2, 1980 (United States)	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays
4	Caddyshack	R	Comedy	1980	July 25, 1980	July 25, 1980 (United States)	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase
…	…	…	…	…	…	…	…	…	…	…	…
7663	More to Life	NaN	Drama	2020	October 23, 2020	October 23, 2020 (United States)	3.1	18.0	Joseph Ebanks	Joseph Ebanks	Shannon Bond
7664	Dream Round	NaN	Comedy	2020	February 7, 2020	February 7, 2020 (United States)	4.7	36.0	Dusty Dukatz	Lisa Huston	Michael Saquella
7665	Saving Mbango	NaN	Drama	2020	April 27, 2020	April 27, 2020 (Cameroon)	5.7	29.0	Nkanya Nkwai	Lynno Lovert	Onyama Laura
7666	It's Just Us	NaN	Drama	2020	October 1, 2020	October 1, 2020 (United States)	NaN	NaN	James Randall	James Randall	Christina Roz
7667	Tee em el	NaN	Horror	2020	August 19, 2020	August 19, 2020 (United States)	5.7	7.0	Pereko Mosia	Pereko Mosia	Siyabonga Mabaso

```
              country      budget        gross                    company  \
0      United Kingdom  19000000.0   46998772.0               Warner Bros.
1       United States   4500000.0   58853106.0           Columbia Pictures
2       United States  18000000.0  538375067.0                   Lucasfilm
3       United States   3500000.0   83453539.0          Paramount Pictures
4       United States   6000000.0   39846344.0              Orion Pictures
…                 …         …           …                         …
7663    United States      7000.0         NaN                       NaN
7664    United States         NaN         NaN  Cactus Blue Entertainment
7665    United States     58750.0         NaN            Embi Productions
7666    United States     15000.0         NaN                       NaN
7667     South Africa         NaN         NaN                 PK 65 Films

      runtime
0       146.0
1       104.0
2       124.0
3        88.0
4        98.0
…        …
7663     90.0
7664     90.0
7665      NaN
7666    120.0
7667    102.0

[7668 rows x 16 columns]
```

[9]:
```
df1.insert(5,'released_place',y['released_place'])
df1
```

[9]:
```
                                                name rating      genre  year  \
0                                        The Shining      R      Drama  1980
1                                   The Blue Lagoon       R  Adventure  1980
2     Star Wars: Episode V - The Empire Strikes Back     PG     Action  1980
3                                          Airplane!     PG     Comedy  1980
4                                       Caddyshack       R     Comedy  1980
…                                                  …    …         …    …
7663                                   More to Life    NaN      Drama  2020
7664                                    Dream Round    NaN     Comedy  2020
7665                                  Saving Mbango    NaN      Drama  2020
7666                                   It's Just Us    NaN      Drama  2020
7667                                      Tee em el    NaN     Horror  2020

          released_date released_place                          released  \
0        June 13, 1980   United States     June 13, 1980 (United States)
```

```
1         July 2, 1980   United States      July 2, 1980 (United States)
2        June 20, 1980   United States     June 20, 1980 (United States)
3         July 2, 1980   United States      July 2, 1980 (United States)
4        July 25, 1980   United States     July 25, 1980 (United States)
…                  …           …                                 …
7663  October 23, 2020   United States  October 23, 2020 (United States)
7664  February 7, 2020   United States  February 7, 2020 (United States)
7665    April 27, 2020        Cameroon         April 27, 2020 (Cameroon)
7666   October 1, 2020   United States   October 1, 2020 (United States)
7667   August 19, 2020   United States   August 19, 2020 (United States)

      score      votes         director                  writer  \
0       8.4   927000.0  Stanley Kubrick             Stephen King
1       5.8    65000.0   Randal Kleiser  Henry De Vere Stacpoole
2       8.7  1200000.0   Irvin Kershner           Leigh Brackett
3       7.7   221000.0     Jim Abrahams             Jim Abrahams
4       7.3   108000.0     Harold Ramis       Brian Doyle-Murray
…     …         …               …                         …
7663    3.1       18.0    Joseph Ebanks             Joseph Ebanks
7664    4.7       36.0     Dusty Dukatz              Lisa Huston
7665    5.7       29.0     Nkanya Nkwai              Lynno Lovert
7666    NaN        NaN    James Randall             James Randall
7667    5.7        7.0     Pereko Mosia              Pereko Mosia

                 star         country      budget        gross  \
0       Jack Nicholson  United Kingdom  19000000.0   46998772.0
1       Brooke Shields   United States   4500000.0   58853106.0
2          Mark Hamill   United States  18000000.0  538375067.0
3          Robert Hays   United States   3500000.0   83453539.0
4          Chevy Chase   United States   6000000.0   39846344.0
…                  …           …          …           …
7663      Shannon Bond   United States      7000.0          NaN
7664  Michael Saquella   United States         NaN          NaN
7665      Onyama Laura   United States     58750.0          NaN
7666     Christina Roz   United States     15000.0          NaN
7667  Siyabonga Mabaso    South Africa         NaN          NaN

                        company  runtime
0                  Warner Bros.    146.0
1             Columbia Pictures    104.0
2                     Lucasfilm    124.0
3            Paramount Pictures     88.0
4                Orion Pictures     98.0
…                           …      …
7663                       NaN     90.0
7664  Cactus Blue Entertainment     90.0
7665           Embi Productions      NaN
```

```
7666                        NaN    120.0
7667                PK 65 Films    102.0

[7668 rows x 17 columns]
```

[10]:
```
df1.drop(['released'],axis=1,inplace=True)
```

[11]:
```
df1.head()
```

[11]:
```
                                       name rating      genre  year  \
0                                The Shining      R      Drama  1980
1                             The Blue Lagoon      R  Adventure  1980
2  Star Wars: Episode V - The Empire Strikes Back     PG     Action  1980
3                                  Airplane!     PG     Comedy  1980
4                                 Caddyshack      R     Comedy  1980

    released_date released_place  score      votes         director  \
0  June 13, 1980   United States    8.4   927000.0  Stanley Kubrick
1   July 2, 1980   United States    5.8    65000.0   Randal Kleiser
2  June 20, 1980   United States    8.7  1200000.0   Irvin Kershner
3   July 2, 1980   United States    7.7   221000.0     Jim Abrahams
4  July 25, 1980   United States    7.3   108000.0     Harold Ramis

                   writer            star         country      budget  \
0             Stephen King  Jack Nicholson  United Kingdom  19000000.0
1  Henry De Vere Stacpoole  Brooke Shields   United States   4500000.0
2          Leigh Brackett     Mark Hamill   United States  18000000.0
3            Jim Abrahams     Robert Hays   United States   3500000.0
4       Brian Doyle-Murray     Chevy Chase   United States   6000000.0

         gross             company  runtime
0   46998772.0        Warner Bros.    146.0
1   58853106.0   Columbia Pictures    104.0
2  538375067.0           Lucasfilm    124.0
3   83453539.0  Paramount Pictures     88.0
4   39846344.0      Orion Pictures     98.0
```

[12]:
```
# Convert Release Date to DateTime
df1['released_date'] = df1['released_date'].astype(str)
df1['released_date'] = pd.to_datetime(df1['released_date'], format='%m/%d/%Y',␣
 ↪errors='coerce')
df1.head()
```

[12]:
```
                                       name rating      genre  year  \
0                                The Shining      R      Drama  1980
1                             The Blue Lagoon      R  Adventure  1980
2  Star Wars: Episode V - The Empire Strikes Back     PG     Action  1980
```

```
3                                    Airplane!     PG     Comedy  1980
4                                   Caddyshack      R     Comedy  1980

  released_date released_place  score      votes         director  \
0           NaT  United States    8.4   927000.0  Stanley Kubrick
1           NaT  United States    5.8    65000.0   Randal Kleiser
2           NaT  United States    8.7  1200000.0   Irvin Kershner
3           NaT  United States    7.7   221000.0     Jim Abrahams
4           NaT  United States    7.3   108000.0     Harold Ramis

                   writer            star         country      budget  \
0             Stephen King  Jack Nicholson  United Kingdom  19000000.0
1  Henry De Vere Stacpoole  Brooke Shields   United States   4500000.0
2           Leigh Brackett     Mark Hamill   United States  18000000.0
3             Jim Abrahams     Robert Hays   United States   3500000.0
4       Brian Doyle-Murray     Chevy Chase   United States   6000000.0

         gross             company  runtime
0   46998772.0        Warner Bros.    146.0
1   58853106.0   Columbia Pictures    104.0
2  538375067.0           Lucasfilm    124.0
3   83453539.0  Paramount Pictures     88.0
4   39846344.0      Orion Pictures     98.0
```

[13]:
```
df1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7668 entries, 0 to 7667
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   name            7668 non-null   object
 1   rating          7591 non-null   object
 2   genre           7668 non-null   object
 3   year            7668 non-null   int64
 4   released_date   0 non-null      datetime64[ns]
 5   released_place  7666 non-null   object
 6   score           7665 non-null   float64
 7   votes           7665 non-null   float64
 8   director        7668 non-null   object
 9   writer          7665 non-null   object
 10  star            7667 non-null   object
 11  country         7665 non-null   object
 12  budget          5497 non-null   float64
 13  gross           7479 non-null   float64
 14  company         7651 non-null   object
 15  runtime         7664 non-null   float64
dtypes: datetime64[ns](1), float64(5), int64(1), object(9)
```

```
memory usage: 958.6+ KB
```

[14]:
```
#Check fofr duplicates
df1.duplicated().value_counts()
```

[14]:
```
False    7668
Name: count, dtype: int64
```

No duplicates

[15]:
```
# Check for null values
df1.isnull().sum().sort_values(ascending= False)
```

[15]:
```
released_date     7668
budget            2171
gross              189
rating              77
company             17
runtime              4
score                3
votes                3
writer               3
country              3
released_place       2
star                 1
name                 0
genre                0
year                 0
director             0
dtype: int64
```

[16]:
```
# Create a pivot table for null value sum
df1.isnull().sum().sort_values(ascending= False)[:-4].reset_index().
  ↪rename(columns={"index":"columns",0:"Null_Values"})
```

[16]:

	columns	Null_Values
0	released_date	7668
1	budget	2171
2	gross	189
3	rating	77
4	company	17
5	runtime	4
6	score	3
7	votes	3
8	writer	3
9	country	3
10	released_place	2
11	star	1

Null values present in 12 columns: 6 object dtype , 5 float64 dtype and 1 is datetime64. Fill

1.For float64 dtype columns with "0".

2.For object dtype columns with "Others".

3.For datetime64 dtype column(released_date) with "0000-00-00".

[17]:
```
# Replacing null values
columns_1 = ['rating','company','writer','country','released_place','star']
for i in columns_1:
    df1[i].fillna("Others",inplace=True)

columns_2 = ['runtime','score','votes']
for j in columns_2:
    df1[j].fillna(0,inplace=True)

columns_3 = ['budget','gross']
for k in columns_3:
    df1[k].fillna(round(np.mean(df1[k])),inplace=True)
```

[18]:
```
df1
```

[18]:

	name	rating	genre	year	released_date	released_place	score	votes	director
0	The Shining	R	Drama	1980	NaT	United States	8.4	927000.0	Stanley Kubrick
1	The Blue Lagoon	R	Adventure	1980	NaT	United States	5.8	65000.0	Randal Kleiser
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	NaT	United States	8.7	1200000.0	Irvin Kershner
3	Airplane!	PG	Comedy	1980	NaT	United States	7.7	221000.0	Jim Abrahams
4	Caddyshack	R	Comedy	1980	NaT	United States	7.3	108000.0	Harold Ramis
…	…	…	…	…	…	…	…	…	…
7663	More to Life	Others	Drama	2020	NaT	United States	3.1	18.0	Joseph Ebanks
7664	Dream Round	Others	Comedy	2020	NaT	United States	4.7	36.0	Dusty Dukatz
7665	Saving Mbango	Others	Drama	2020	NaT	Cameroon	5.7	29.0	Nkanya Nkwai
7666	It's Just Us	Others	Drama	2020	NaT	United States	0.0	0.0	James Randall
7667	Tee em el	Others	Horror	2020					

```
7667           NaT  United States    5.7        7.0     Pereko Mosia

                      writer              star         country      budget  \
0                Stephen King    Jack Nicholson  United Kingdom  19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States   4500000.0
2               Leigh Brackett       Mark Hamill   United States  18000000.0
3                 Jim Abrahams       Robert Hays   United States   3500000.0
4           Brian Doyle-Murray       Chevy Chase   United States   6000000.0
…                          …                 …              …         …
7663             Joseph Ebanks      Shannon Bond   United States      7000.0
7664               Lisa Huston  Michael Saquella   United States  35589876.0
7665               Lynno Lovert      Onyama Laura   United States     58750.0
7666             James Randall     Christina Roz   United States     15000.0
7667              Pereko Mosia  Siyabonga Mabaso    South Africa  35589876.0

            gross                    company  runtime
0      46998772.0               Warner Bros.    146.0
1      58853106.0          Columbia Pictures    104.0
2     538375067.0                   Lucasfilm    124.0
3      83453539.0         Paramount Pictures     88.0
4      39846344.0             Orion Pictures     98.0
…             …                          …      …
7663   78500541.0                     Others     90.0
7664   78500541.0  Cactus Blue Entertainment     90.0
7665   78500541.0           Embi Productions      0.0
7666   78500541.0                     Others    120.0
7667   78500541.0                PK 65 Films    102.0

[7668 rows x 16 columns]
```

[19]:
```
df1 = df1.drop('released_date', axis=1)
df1
```

[19]:
```
                                          name  rating      genre  year  \
0                                   The Shining       R      Drama  1980
1                                The Blue Lagoon       R  Adventure  1980
2     Star Wars: Episode V - The Empire Strikes Back      PG     Action  1980
3                                     Airplane!      PG     Comedy  1980
4                                   Caddyshack       R     Comedy  1980
…                                             …     …          …     …
7663                             More to Life  Others      Drama  2020
7664                              Dream Round  Others     Comedy  2020
7665                            Saving Mbango  Others      Drama  2020
7666                             It's Just Us  Others      Drama  2020
7667                                Tee em el  Others     Horror  2020

     released_place  score      votes         director  \
```

```
0     United States    8.4   927000.0  Stanley Kubrick
1     United States    5.8    65000.0   Randal Kleiser
2     United States    8.7  1200000.0   Irvin Kershner
3     United States    7.7   221000.0     Jim Abrahams
4     United States    7.3   108000.0     Harold Ramis
…               …    …        …               …
7663  United States    3.1       18.0    Joseph Ebanks
7664  United States    4.7       36.0     Dusty Dukatz
7665       Cameroon    5.7       29.0     Nkanya Nkwai
7666  United States    0.0        0.0    James Randall
7667  United States    5.7        7.0     Pereko Mosia

                        writer              star         country      budget  \
0                Stephen King    Jack Nicholson  United Kingdom  19000000.0
1     Henry De Vere Stacpoole    Brooke Shields   United States   4500000.0
2               Leigh Brackett       Mark Hamill   United States  18000000.0
3                 Jim Abrahams       Robert Hays   United States   3500000.0
4           Brian Doyle-Murray       Chevy Chase   United States   6000000.0
…                          …                 …               …          …
7663             Joseph Ebanks      Shannon Bond   United States      7000.0
7664               Lisa Huston  Michael Saquella   United States  35589876.0
7665               Lynno Lovert      Onyama Laura   United States     58750.0
7666              James Randall     Christina Roz   United States     15000.0
7667               Pereko Mosia  Siyabonga Mabaso    South Africa  35589876.0

            gross                    company  runtime
0      46998772.0               Warner Bros.    146.0
1      58853106.0          Columbia Pictures    104.0
2     538375067.0                   Lucasfilm    124.0
3      83453539.0         Paramount Pictures     88.0
4      39846344.0             Orion Pictures     98.0
…             …                          …      …
7663   78500541.0                     Others     90.0
7664   78500541.0  Cactus Blue Entertainment     90.0
7665   78500541.0           Embi Productions      0.0
7666   78500541.0                     Others    120.0
7667   78500541.0               PK 65 Films    102.0

[7668 rows x 15 columns]
```

[20]:
```
df1.isna().sum().sort_values(ascending=False)[:-4].reset_index().
↪rename(columns={"index":"columns",0:"Null_Values"})
```

[20]:
```
           columns  Null_Values
0             name            0
1           rating            0
2            genre            0
```

```
3              year            0
4   released_place             0
5             score            0
6             votes            0
7          director            0
8            writer            0
9              star            0
10          country            0
```

[21]:
```
df1.shape
```

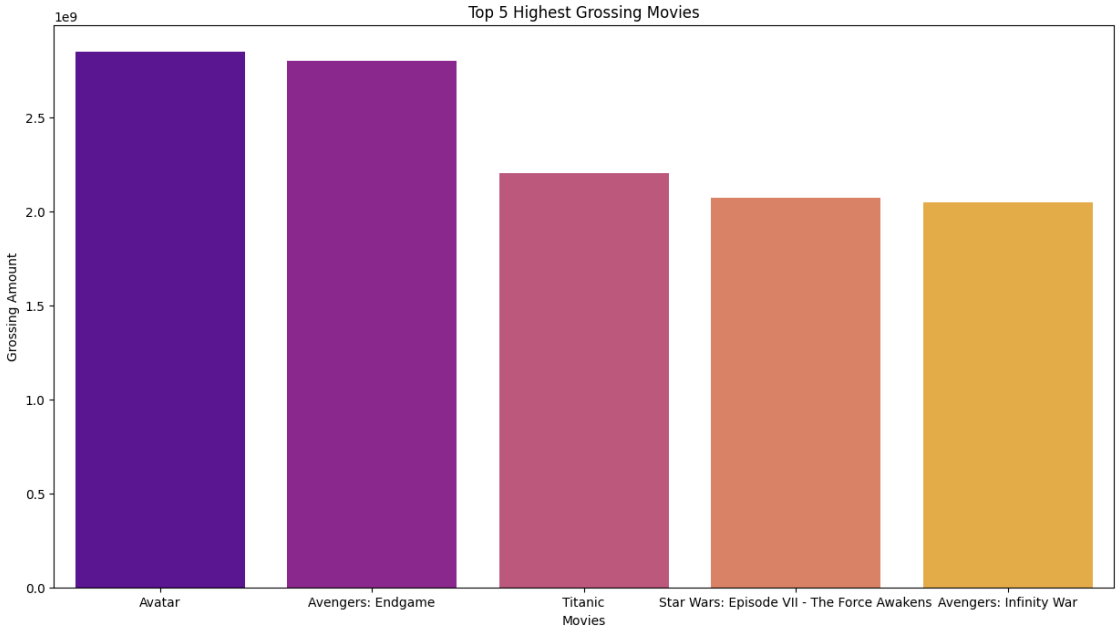
[21]:
```
(7668, 15)
```

[22]:
```
# Save Cleaned and Manipulated data to file
df1.to_csv("Movie_industry.csv")
```

0.0.6 Analysis and Visualization of Data

[23]:
```
# Top 5 Grossing Movies
plt.figure(figsize=(15,8))
#sns.barplot(x='name',y='gross',data=df1.
 ↪sort_values(by='gross',ascending=False).head(),palette='plasma')
sns.barplot(x='name', y='gross', hue='name', data=df1.sort_values(by='gross',␣
 ↪ascending=False).head(), palette='plasma', legend=False)
plt.xlabel('Movies')
plt.ylabel('Grossing Amount')
plt.title('Top 5 Highest Grossing Movies')
```

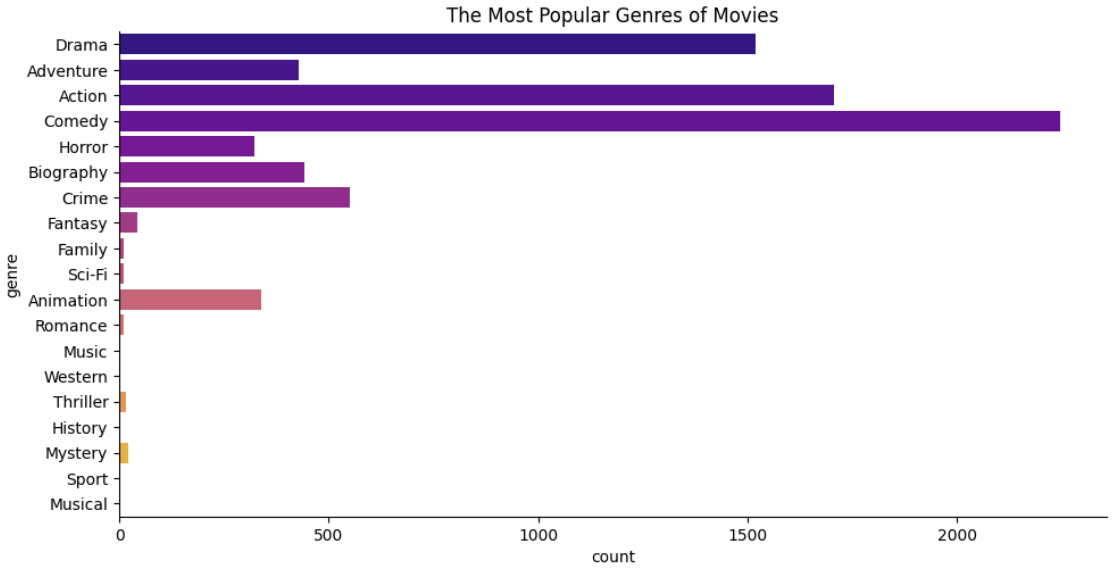
[23]:
```
Text(0.5, 1.0, 'Top 5 Highest Grossing Movies')
```



Avatar is highest grossing movie

[24]:
```
# Popular genres of movies
# Plot bar chart with total movies of each genre
plt.figure(figsize=(15,10))
sns.catplot(y='genre', kind='count', hue= 'genre',legend=False, data=df1,␣
 ↪aspect=2, palette="plasma");
plt.title('The Most Popular Genres of Movies')
```

[24]:
```
Text(0.5, 1.0, 'The Most Popular Genres of Movies')

<Figure size 1500x1000 with 0 Axes>
```

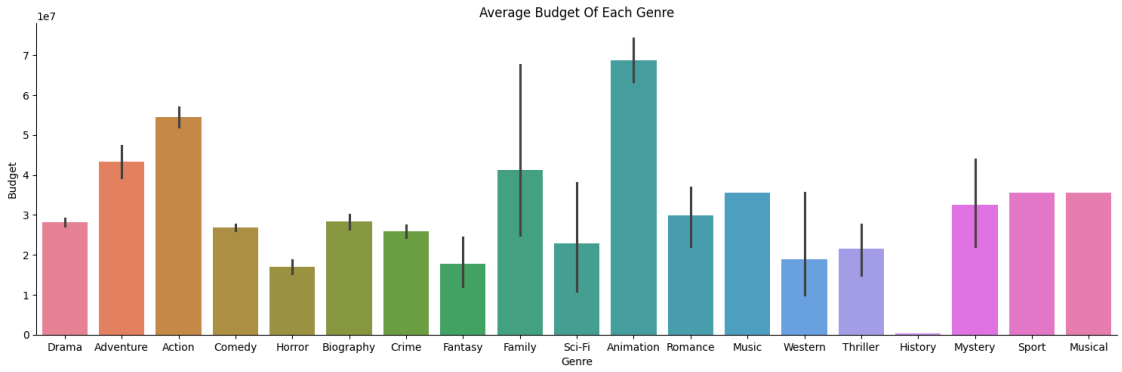


Most Popular genre is Comedy

[25]:
```
# Plot bar chart with average budget of each genre
plt.tight_layout()
plt.figure(figsize=(10,10))
sns.catplot(x='genre', y='budget', kind='bar', data=df1,␣
 ↪aspect=3,hue='genre',legend=False)
plt.xlabel('Genre')
plt.ylabel('Budget')
plt.title('Average Budget Of Each Genre')
```

[25]:
```
Text(0.5, 1.0, 'Average Budget Of Each Genre')

<Figure size 640x480 with 0 Axes>
```

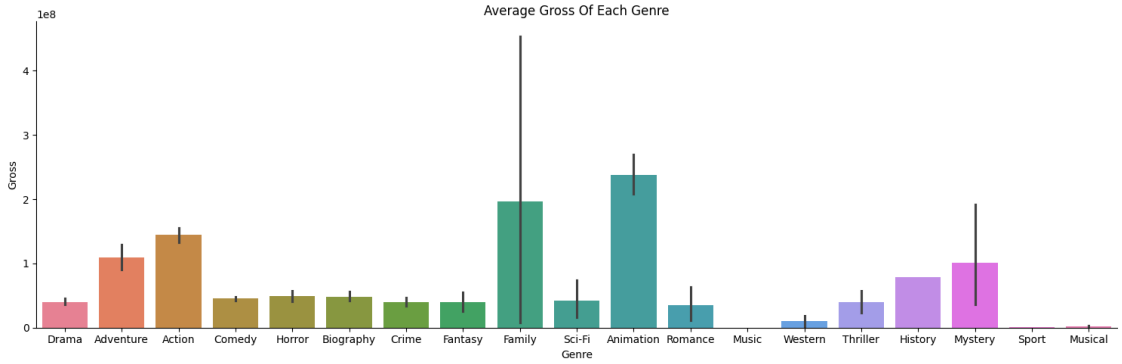
<Figure size 1000x1000 with 0 Axes>



Animated movies(Genre: animation) have the highest average budget

[26]:
```
# Plot bar chart with average gross of each genre
plt.figure(figsize=(10,8))
sns.catplot(x='genre', y='gross', kind='bar', data=df1, aspect=3,hue='genre',␣
 ↪legend= False)
plt.xlabel('Genre')
plt.ylabel('Gross')
plt.title('Average Gross Of Each Genre')
plt.tight_layout()
plt.show()
```

<Figure size 1000x800 with 0 Axes>



Animated movies(Genre:animation)have highest gross of each genre

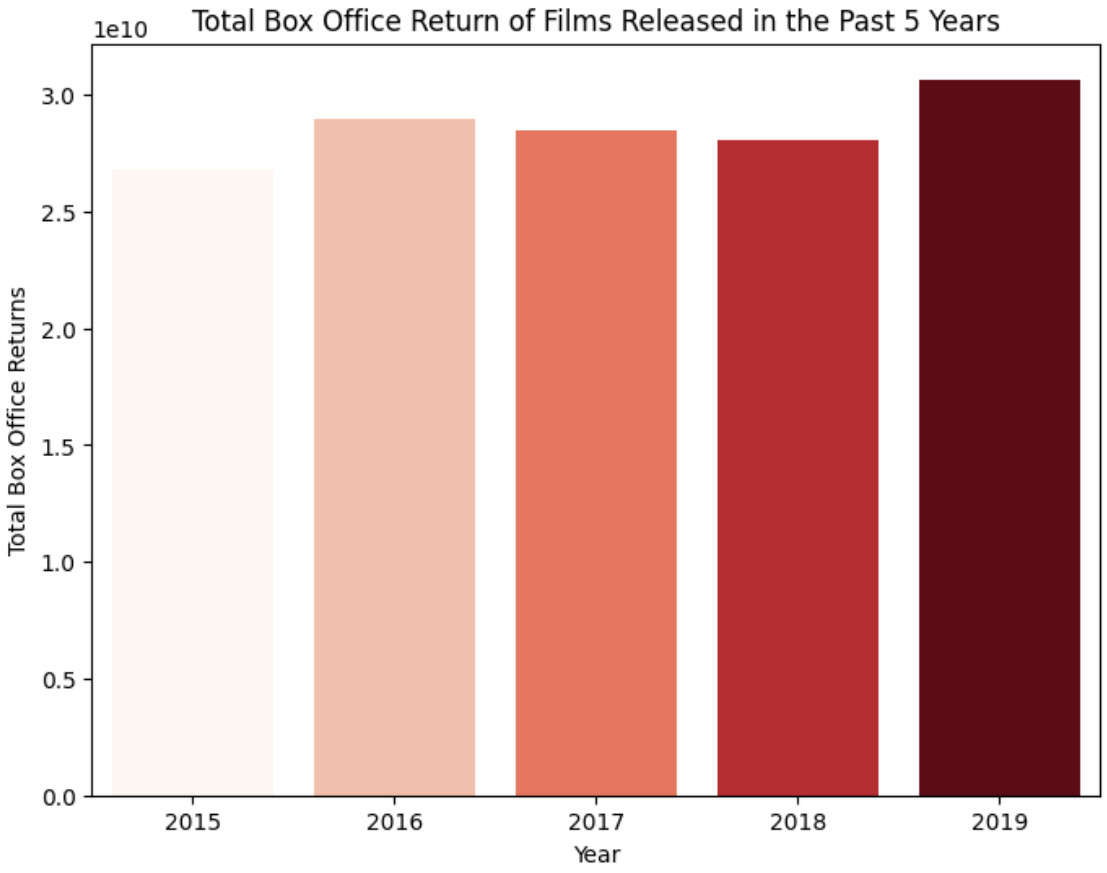
[27]:
```
# Total box office return of films released in the past 5 years
#Analysis
```

```
data = df1.groupby('year')['gross'].agg({'sum'}).
 ↪sort_values(by='year',ascending = False)[1:6].reset_index().
 ↪rename(columns={'sum':'total_box_office_returns'})
data
```

[27]:

	year	total_box_office_returns
0	2019	3.065366e+10
1	2018	2.808529e+10
2	2017	2.848525e+10
3	2016	2.893884e+10
4	2015	2.682451e+10

[28]:
```
# Total box office return of films released in the past 5 years
# Visualization

plt.figure(figsize=(8,6))
sns.barplot(x='year', y='total_box_office_returns', data=data, hue='year',␣
 ↪palette='Reds', legend=False)
plt.xlabel('Year')
plt.ylabel('Total Box Office Returns')
plt.title("Total Box Office Return of Films Released in the Past 5 Years")
plt.show()
```



2019 has the highest of Total Box Office Return of Films Released From(2015-2019)

[29]:
```
# Total profits of past 10 years in movie industry
# Analysis

# Profit Formula "Total_Sales-Total_Expenses".
df1['profit'] = df1['gross']-df1['budget']
df1.head()
```

[29]:

	name	rating	genre	year	released_place	score	votes	director	writer	star	country	budget	gross	company	runtime	profit
0	The Shining	R	Drama	1980	United States	8.4	927000.0	Stanley Kubrick	Stephen King	Jack Nicholson	United Kingdom	19000000.0	46998772.0	Warner Bros.	146.0	27998772.0
1	The Blue Lagoon	R	Adventure	1980	United States	5.8	65000.0	Randal Kleiser	Henry De Vere Stacpoole	Brooke Shields	United States	4500000.0	58853106.0	Columbia Pictures	104.0	54353106.0
2	Star Wars: Episode V - The Empire Strikes Back	PG	Action	1980	United States	8.7	1200000.0	Irvin Kershner	Leigh Brackett	Mark Hamill	United States	18000000.0	538375067.0	Lucasfilm	124.0	520375067.0
3	Airplane!	PG	Comedy	1980	United States	7.7	221000.0	Jim Abrahams	Jim Abrahams	Robert Hays	United States	3500000.0	83453539.0	Paramount Pictures	88.0	79953539.0
4	Caddyshack	R	Comedy	1980	United States	7.3	108000.0	Harold Ramis	Brian Doyle-Murray	Chevy Chase	United States	6000000.0	39846344.0	Orion Pictures	98.0	33846344.0

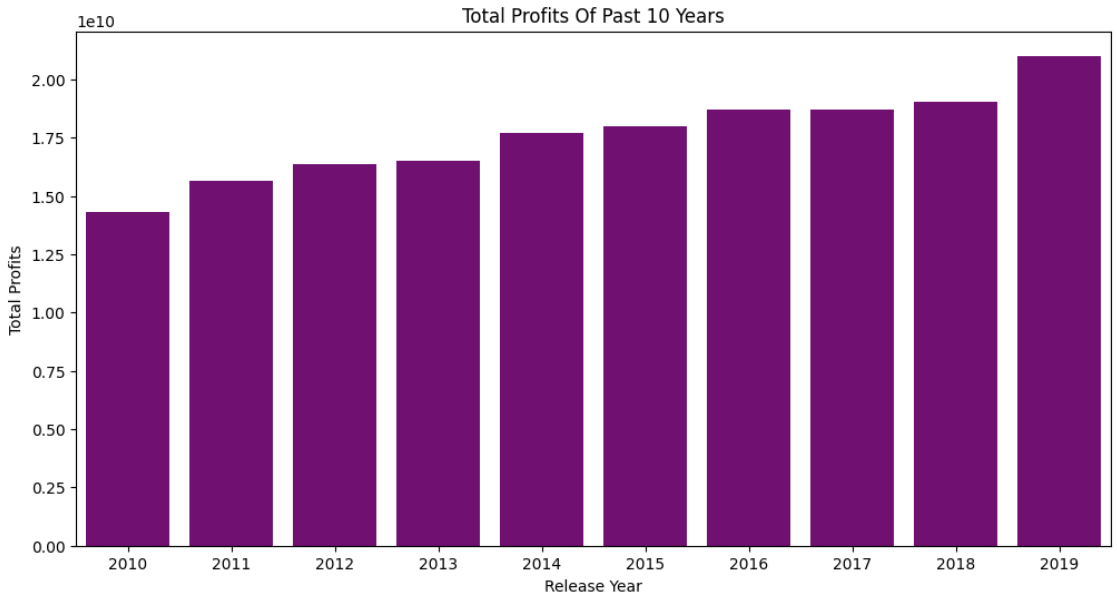
[30]:
```
# Calculate total profit of each year
df_profit = df1.groupby('year')['profit'].sum().reset_index().sort_values(by␣
 ↪='year',ascending=False).rename(columns={'profit':'total_profits'})[1:11]
df_profit
```

[30]:

	year	total_profits
39	2019	2.100556e+10
38	2018	1.904369e+10

```
37  2017    1.869103e+10
36  2016    1.869316e+10
35  2015    1.796502e+10
34  2014    1.772036e+10
33  2013    1.650617e+10
32  2012    1.638207e+10
31  2011    1.562386e+10
30  2010    1.432961e+10
```

[31]:
```
#  Total profits of Past 10 Years.
# Visualiztion

plt.figure(figsize=(12,6))
sns.barplot(x='year', y='total_profits', data=df_profit,color='purple')
plt.xlabel('Release Year')
plt.ylabel('Total Profits')
plt.title('Total Profits Of Past 10 Years')
plt.show()
```

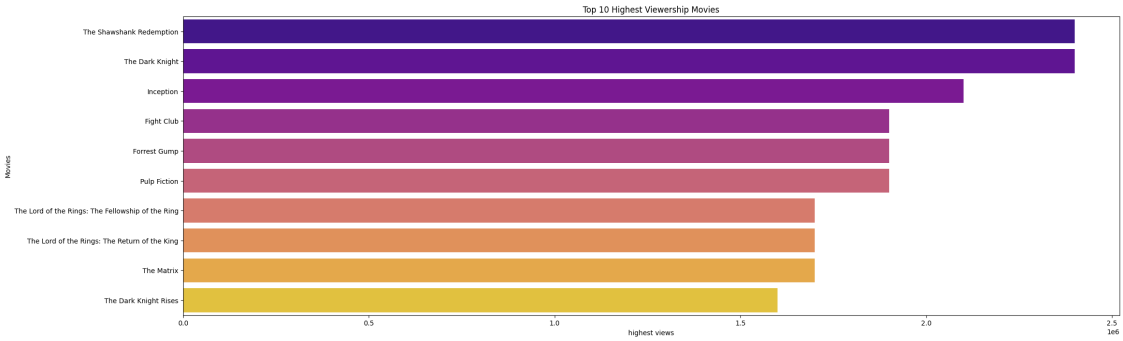


2019 have the highest Total Profits of Past 10 Years(2010-2019)

[32]:
```
# Top 10 movies have the highest viewership
# Visualization
plt.figure(figsize=(25,8))
sns.barplot(y='name',x='votes',data=df1.sort_values(by='votes',ascending=False).
 ↪head(10),palette='plasma',hue='name',legend=False)
plt.ylabel('Movies')
```

```
plt.xlabel('highest views')
plt.title('Top 10 Highest Viewership Movies')
```

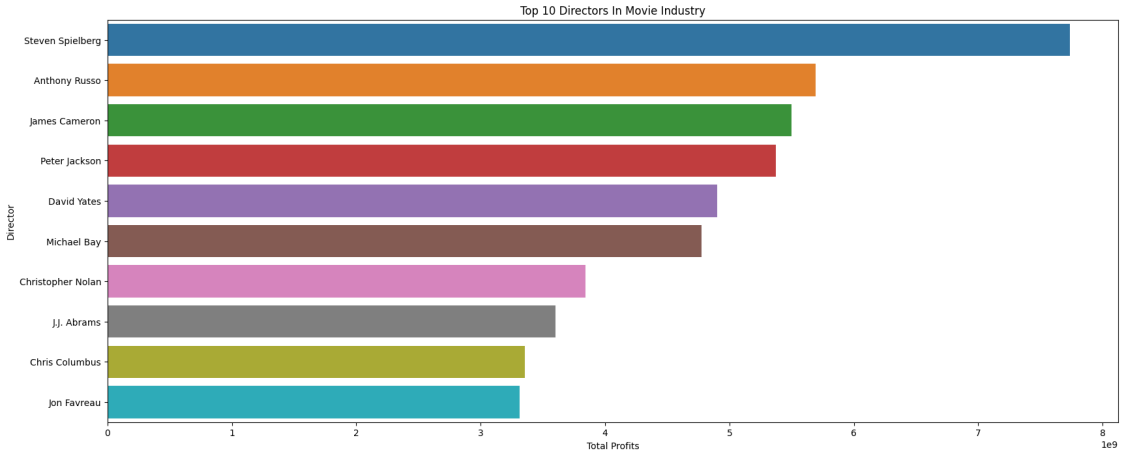
[32]: Text(0.5, 1.0, 'Top 10 Highest Viewership Movies')



The Shawshank Redemption and The Dark Knight have the highest viewership out of top 10 movies have the highest viewership

[33]:
```
# Top 10 Directors In Movie Industry
data = df1.groupby('director')['profit'].sum().reset_index().
 ↪rename(columns={'profit':'total_profit'}).
 ↪sort_values(by='total_profit',ascending=False)[:10]
plt.figure(figsize=(20,8))
sns.barplot(y='director',x='total_profit',data=data,hue='director',␣
 ↪legend=False,)
sns.color_palette()
plt.ylabel('Director')
plt.xlabel('Total Profits')
plt.title('Top 10 Directors In Movie Industry')
```

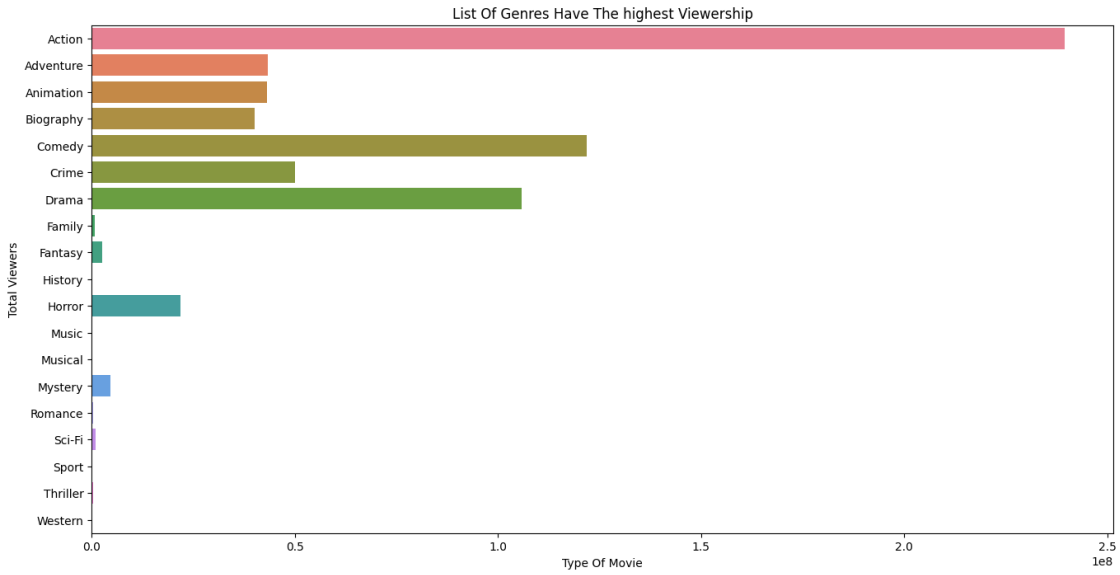
[33]: Text(0.5, 1.0, 'Top 10 Directors In Movie Industry')



Steven Spielberg is the top out of top 10 Directors in Movie Industry

[34]:
```
# List Of Genres Have The highest Viewership

data = df1.groupby('genre')['votes'].agg({'sum'}).reset_index().
 ↪rename(columns={'sum':'total_views'})
plt.figure(figsize=(16,8))
sns.barplot(x='total_views',y='genre',data=data,hue='genre',legend=False)
sns.color_palette("husl", 8)
plt.xlabel('Type Of Movie')
plt.ylabel('Total Viewers')
plt.title('List Of Genres Have The highest Viewership')
```

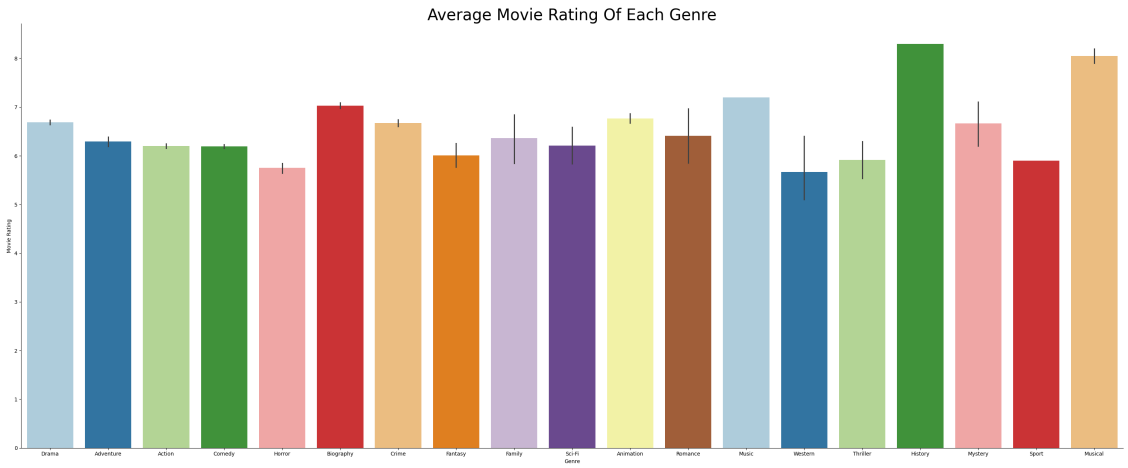
[34]: Text(0.5, 1.0, 'List Of Genres Have The highest Viewership')



Action movies (Genre:Action) is the top genres which have highest Viewership

[35]:
```
# Average Movie Rating Of Each Genre
sns.catplot(x='genre', y='score', kind='bar', data=df1, height=12, aspect=2.5,␣
 ↪palette='Paired',hue='genre',legend=False)
#sns.color_palette("Paired")
plt.xlabel('Genre')
plt.ylabel('Movie Rating')
plt.title('Average Movie Rating Of Each Genre' ,size=30)
```

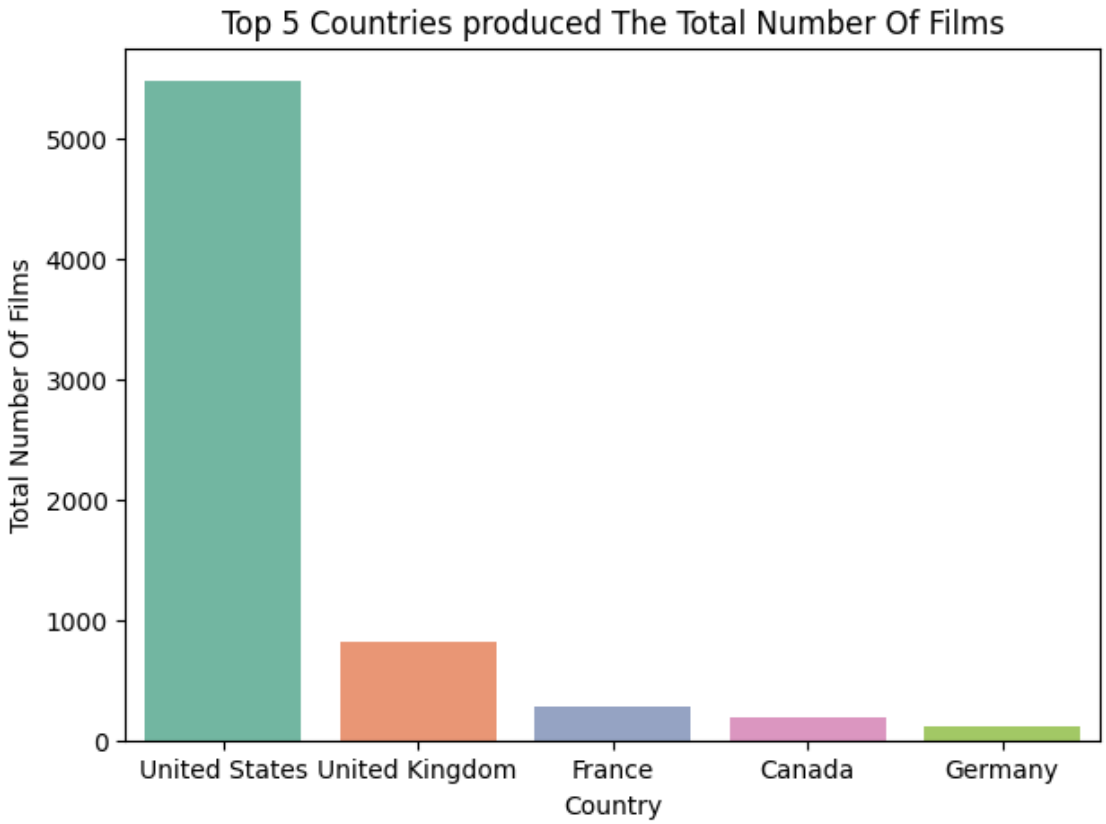
[35]: Text(0.5, 1.0, 'Average Movie Rating Of Each Genre')



History has the highest of Average Movie Rating Of each Genre

[36]:
```
# Top 5 Countries produced The Total Number Of Films
data = df1.groupby('country')['name'].agg({'count'}).reset_index().
 ↪rename(columns={'count':'total_number_of_films'}).
 ↪sort_values(by='total_number_of_films',ascending = False)[:5]
plt.figure(figsize=(7,5))
sns.
 ↪barplot(x='country',y='total_number_of_films',data=data,hue='country',palette="Set2",legend
plt.xlabel('Country')
plt.ylabel('Total Number Of Films')
plt.title('Top 5 Countries produced The Total Number Of Films')
```

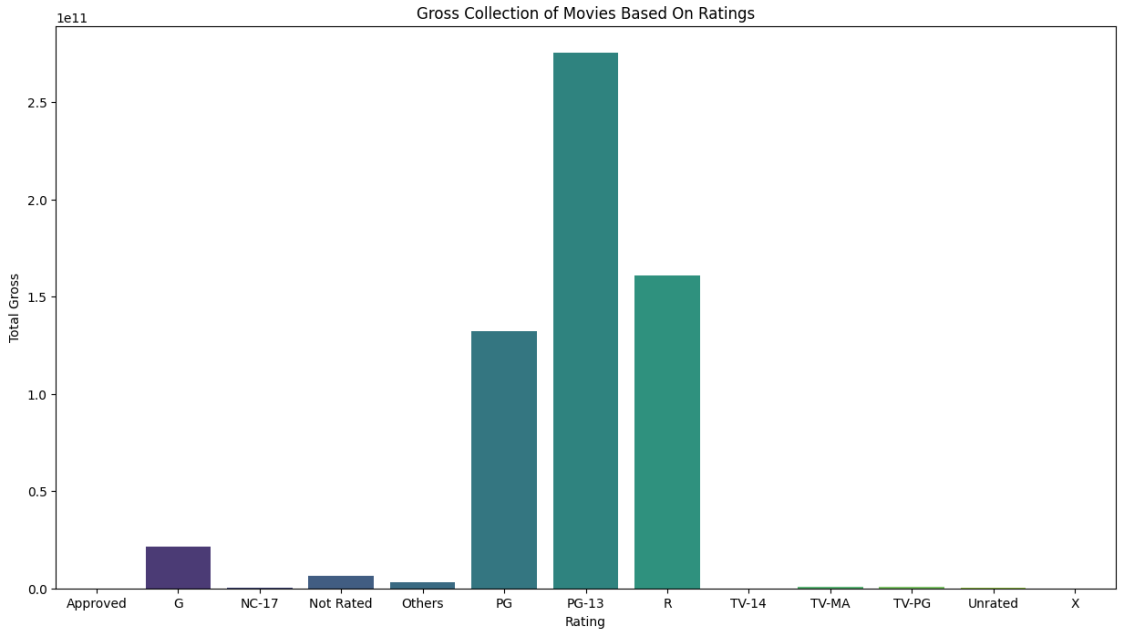
[36]:
```
Text(0.5, 1.0, 'Top 5 Countries produced The Total Number Of Films')
```



United States produce the more movies than any other Country

[37]:
```
# Gross Collection of Movies Based On Ratings
data = df1.groupby('rating')['gross'].sum().reset_index().
 ↪rename(columns={'gross':'total_gross'})
plt.figure(figsize=(15,8))
sns.
 ↪barplot(x='rating',y='total_gross',data=data,hue='rating',palette='viridis',legend=␣
 ↪False)
plt.xlabel('Rating')
plt.ylabel('Total Gross')
plt.title('Gross Collection of Movies Based On Ratings')
```

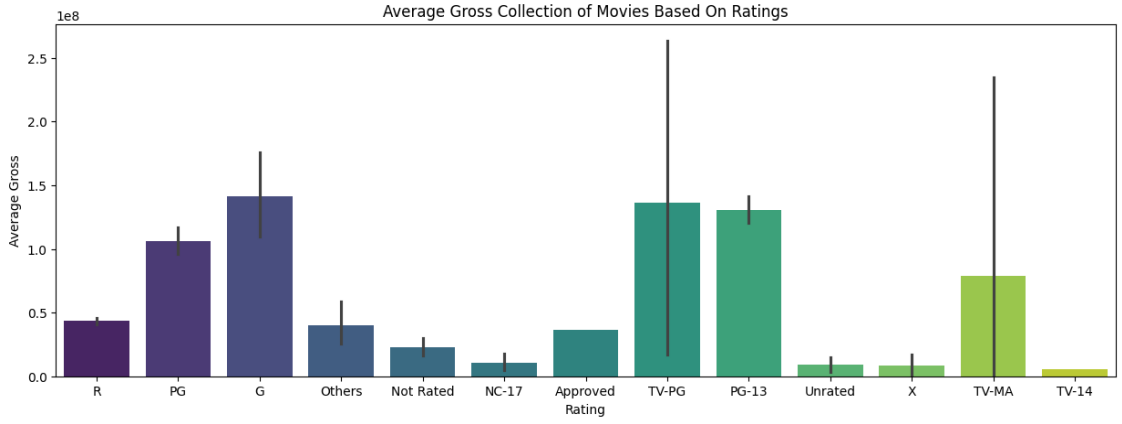
[37]:
```
Text(0.5, 1.0, 'Gross Collection of Movies Based On Ratings')
```



PG-13 movies have highest Gross Collection of movies based on Ratings

[38]:
```
# Average Gross Collections Of Movies Based On Ratings.

plt.figure(figsize=(15,5))
sns.
  ↪barplot(x='rating',y='gross',data=df1,hue='rating',palette='viridis',legend=False)
plt.xlabel('Rating')
plt.ylabel('Average Gross')
plt.title('Average Gross Collection of Movies Based On Ratings')
```

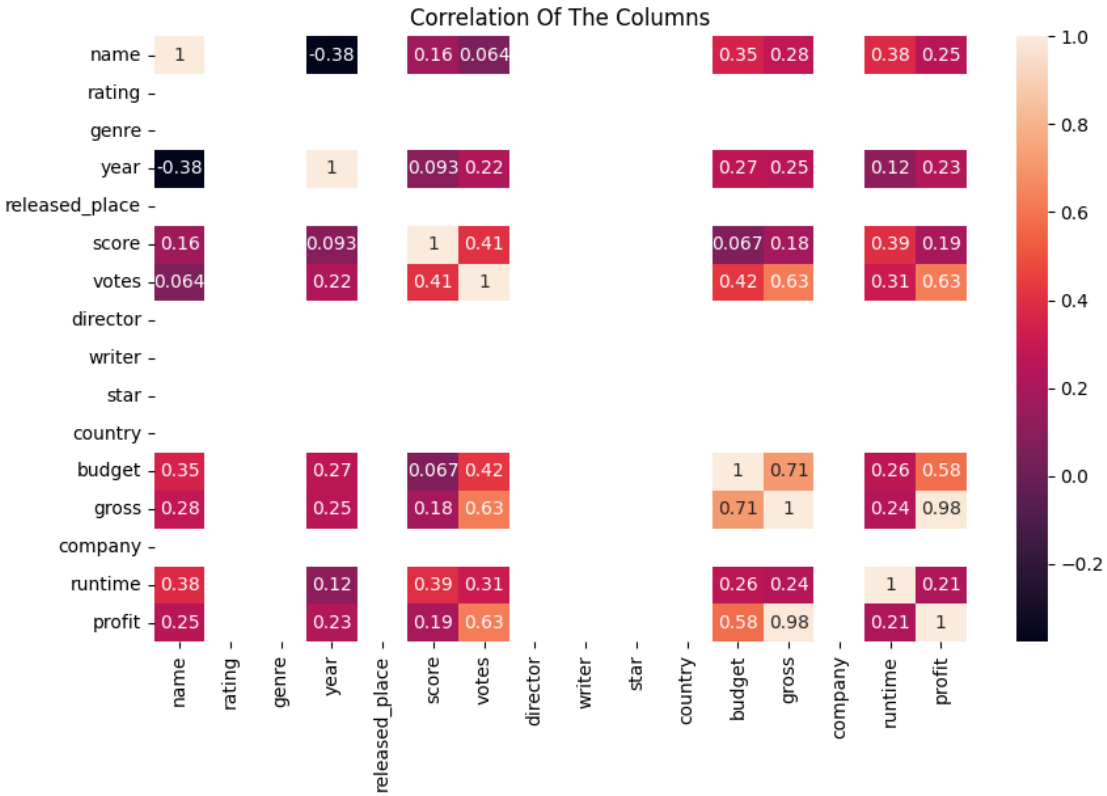
[38]:
```
Text(0.5, 1.0, 'Average Gross Collection of Movies Based On Ratings')
```



G-rated movies have highest of Average Gross Collections of movies based on Ratings.

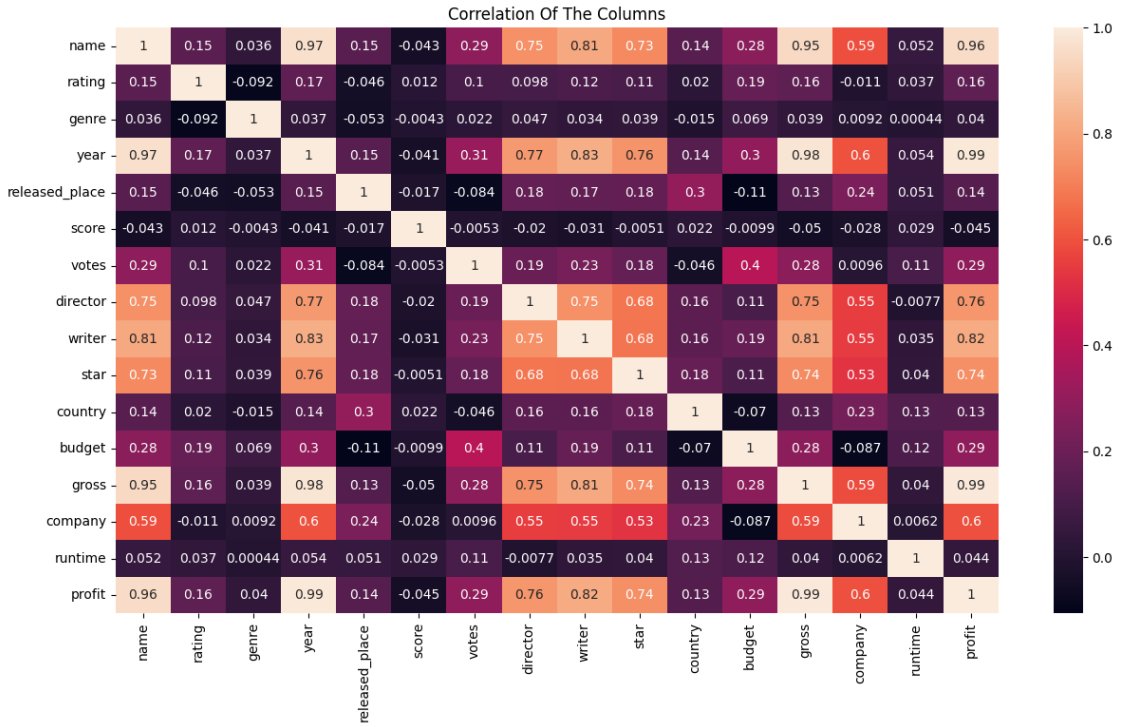
0.0.7 Correlation between movies

[39]:
```
# Correlation between movies
#df1= df1.
 ↪drop('rating','genre','released_place','director','writer','star','company',␣
 ↪axis=1)
df1_numeric = df1.apply(pd.to_numeric, errors='coerce')

plt.figure(figsize=(10, 6))
sns.heatmap(df1_numeric.corr(), annot=True)
sns.color_palette('crest')
plt.title('Correlation Of The Columns')
plt.show()

#plt.figure(figsize=(8,6))
#sns.heatmap(df1.corr(),annot=True)
#plt.title('Correlation Of The Columns')
```



[40]:
```
# Using factorize for correaltion (-assigns a random numeric value for each unique categorical value)

plt.figure(figsize=(15,8))
sns.heatmap(df1.apply(lambda x: x.factorize()[0]).corr(method='pearson'),annot=True)
plt.title('Correlation Of The Columns')
```

[40]: Text(0.5, 1.0, 'Correlation Of The Columns')



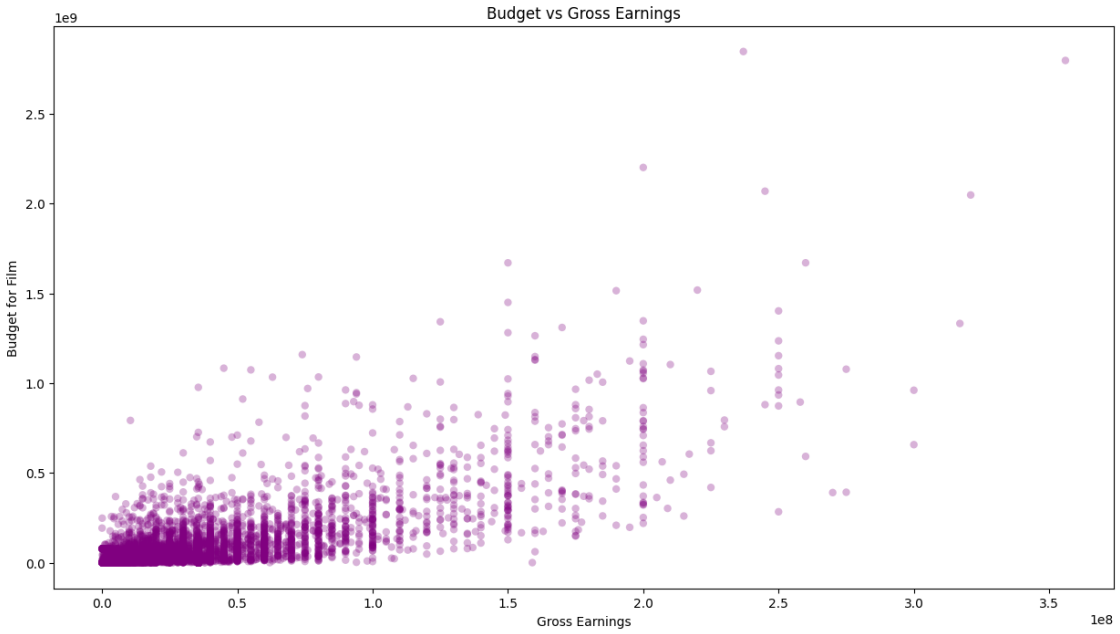
1.A positive correlation between budget and gross earnings in the film industry suggests that as budget of a film increases, so too does the gross earnings of the film making a larger profit.

2.Positive correlation between budget and votes in the film industry suggests that higher budgets tend to lead to higher votes for a film.

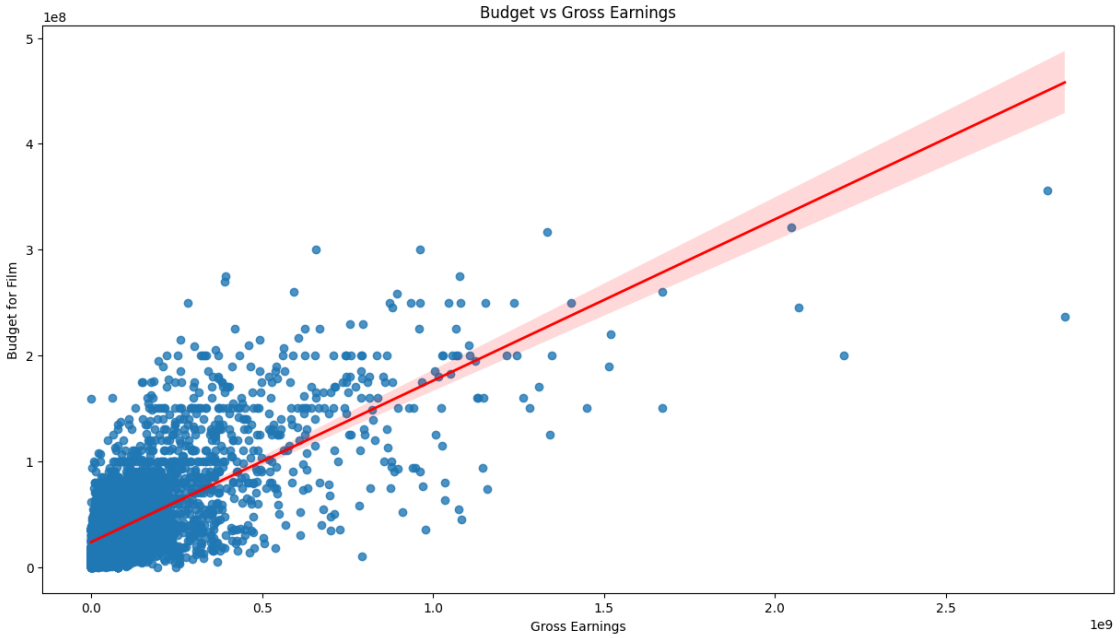
3.A highly positive correlation between director and gross in the film industry suggests that the more successful a director is, the higher the gross of the film will be.

4.Highly positive correlation between star hero and profit in film industry indicates that when a star hero is featured in a film, it is likely to result in higher box office earnings as they have more fan base.

[41]:
```
# Budget vs Gross Earnings

plt.figure(figsize=(15,8))
plt.scatter(x=df1['budget'], y=df1['gross'], alpha=0.
 ↪3,color='purple',edgecolor='none')
plt.title('Budget vs Gross Earnings')
plt.xlabel('Gross Earnings')
plt.ylabel('Budget for Film')
plt.show()
```



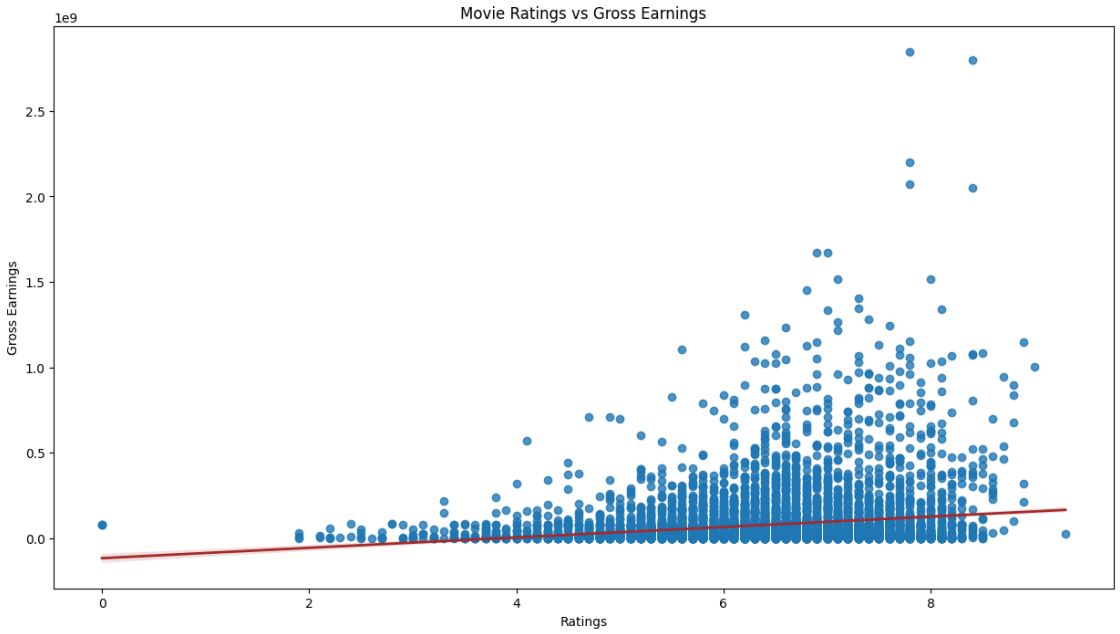
[42]:
```
plt.figure(figsize=(15,8))
sns.regplot(x="gross", y="budget", data=df1,line_kws={'lw':2,'color':'Red'})
plt.title('Budget vs Gross Earnings')
plt.xlabel('Gross Earnings')
plt.ylabel('Budget for Film')
plt.show()
```



The cluster in scatter is positively related to budget and gross.

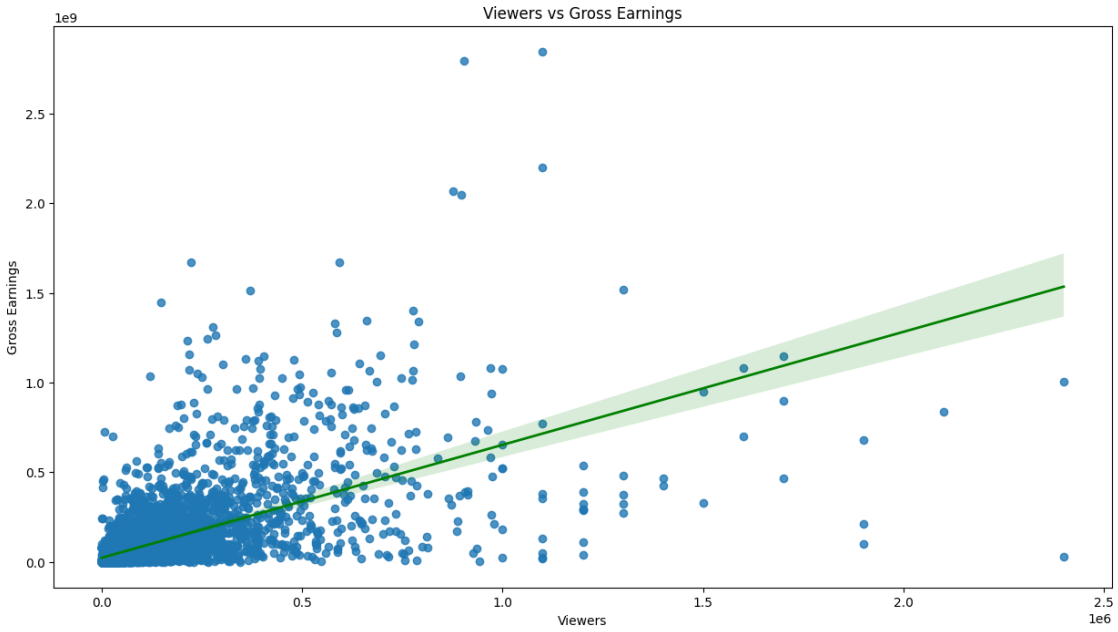
The huge budget implies that increase in quality of the film, screenplay and graphics which attracts more audience to the theater and automatically increase gross earnings.

[43]:
```
# Movie Ratings vs Gross Earnings

plt.figure(figsize=(15,8))
sns.regplot(x="score", y="gross", data=df1,line_kws={'lw':2,'color':'Brown'})
plt.title('Movie Ratings vs Gross Earnings')
plt.xlabel('Ratings')
plt.ylabel('Gross Earnings')
plt.show()
```



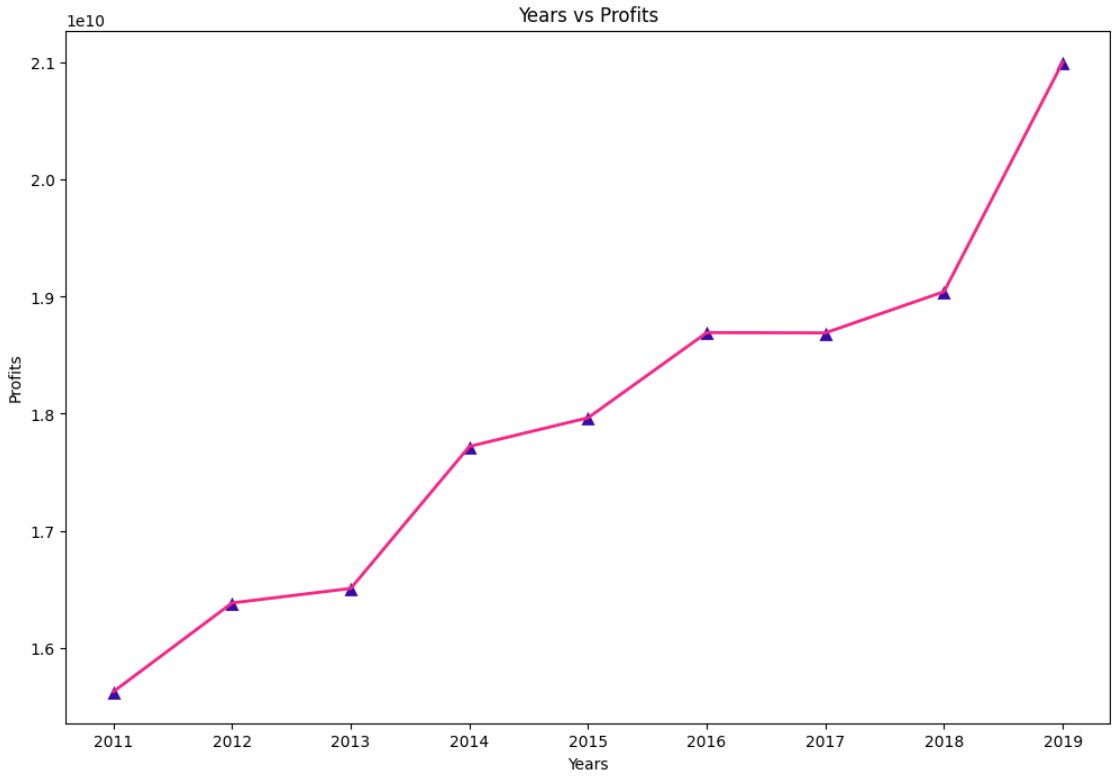
Movies with 6 and above rating score generates more revenue

[44]:
```
# Viewers vs Gross Earnings

plt.figure(figsize=(15,8))
sns.regplot(x="votes", y="gross", data=df1,line_kws={'lw':2,'color':'Green'})
plt.title('Viewers vs Gross Earnings')
plt.xlabel('Viewers')
plt.ylabel('Gross Earnings')
plt.show()
```



The cluster positively correlated to Viewers and Gross Earnigs implies that highly viewed movies tend to generate more revenue than other movies because they are appreciated by audiences.

[45]:
```
# Years vs Profits
data =df1.groupby(['year'])[['profit']].sum()[-10:-1].reset_index()
plt.figure(figsize=(12,8))
sns.lineplot(x=data['year'], y=data['profit'],linewidth=2,color='#F72585')
sns.scatterplot(x=data['year'], y=data['profit'], color='#3A0CA3', s=100,␣
↪marker='^')
plt.title('Years vs Profits')
plt.xlabel('Years')
plt.ylabel('Profits')
plt.show()
```



The graph suggests the more number of movies released in years have inscreased the gross profits.

0.0.8 Conclusion:

The movie and its profits depends on various factors such as budget, star cast, director, stoyline, viewership, rating and votes after release. So we can conclude that if the movie have higher budget, best star cast, best director and best story gives higher viewership, better rating and more votes.

The slowdown could be because of lack of any factors of movie such as less budget, not a good star cast, not best directed by director which recieves less viewership,