TMDb movie data project

Questions asked :

1) who are the top 3 directors have the most successful movies and and what is the number of their movies?

2) what is the percentage of their success ?

3) what is the movie with the highest and lowest budget based on a reference year dollars (2010) ?

4) How is the total revenue of movies over years based on a reference year dollars (2010) ?

5) How is the total budget of movies over years based on a reference year dollars (2010) ?

6) How is the total profit of movies over years based on a reference year dollars (2010) ?

7) How many movies are released in each year over years?

8) How is the runtime of movies over years ?

9) what is the longest and shortest movie ?

10) what is the number of movies that have homepage ?

11) what is the percentage of movies that have homepage ?

12) what is the effect of having homepage on popularity of movie ?

13) What are the top 10 popular movies ?

14) What are the top 10 rated movies ?

15) what are the most 10 profitable movies based on a reference year dollars (2010)?

16) who are the top 10 directors based on the number of released movies ?

17) what is the effect of popularity on profit ?

18) what is the effect of runtime on average voting (rating) ?

19) what is the effect of budget on profit ?

20) what is the top 10 production companies based on number of released movies ?

data wrangling

fill na values:

after getting the data , loading it and assessing it , I found out some missing data in some columns

Categorical Columns with missing data :

[imdb_id , cast , homepage , director , tagline , keywords , overview , genres ,production_companies]

number of missing values in imdb_id: 10

number of missing values in cast: 76

number of missing values in homepage: 7930

number of missing values in director: 44

number of missing values in tagline: 2824

number of missing values in keywords: 1493

number of missing values in overview: 4

number of missing values in genres: 23

number of missing values in production_companies: 1030

I filled missing data with empty string

remove duplicates:

Every movie has a unique id , so I removed duplicated rows that have the same movie id

It was only one row

handling the outlier of numeric columns

numeric columns that have zeros :

[budget, revenue , runtime , budget_adj , revenue_adj]

budget has 5696 zeros

revenue has 6016 zeros

runtime has 31 zeros

budget_adj has 5696 zeros

revenue_adj has 6016 zeros

I filled the zeros with the median of the column in the releasing year of that movie

Description of investigation:

To investigate these questions:

i use pandas functions to :

Select rows and columns using iloc and loc functions

Select some rows based on filtering the dataframe based one conditions

Groub by some columns and use aggregate functions to get number like sum , mean , max ,min

Get the index of the max and minimum of a columns using idxmin() , idxmax()

And use matplotlib to draw visualizations like :

Bar chart

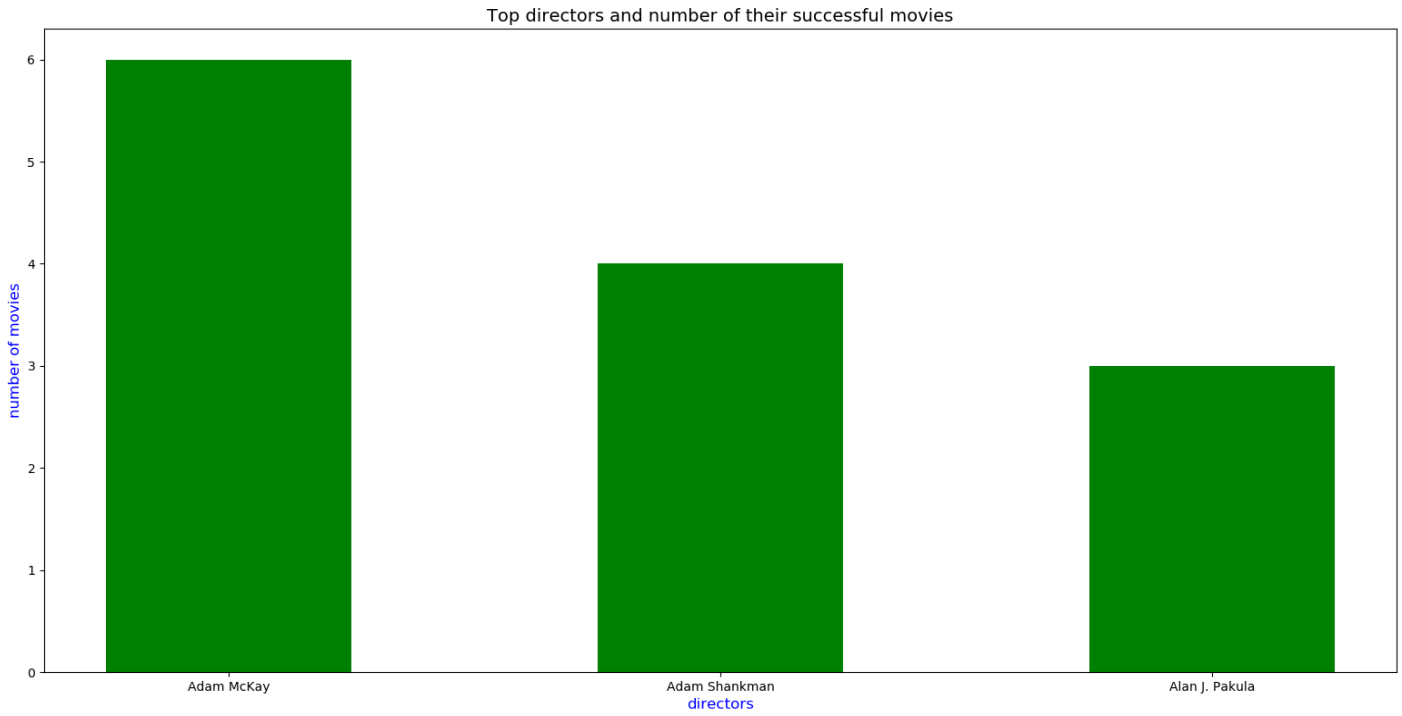
Pie chart

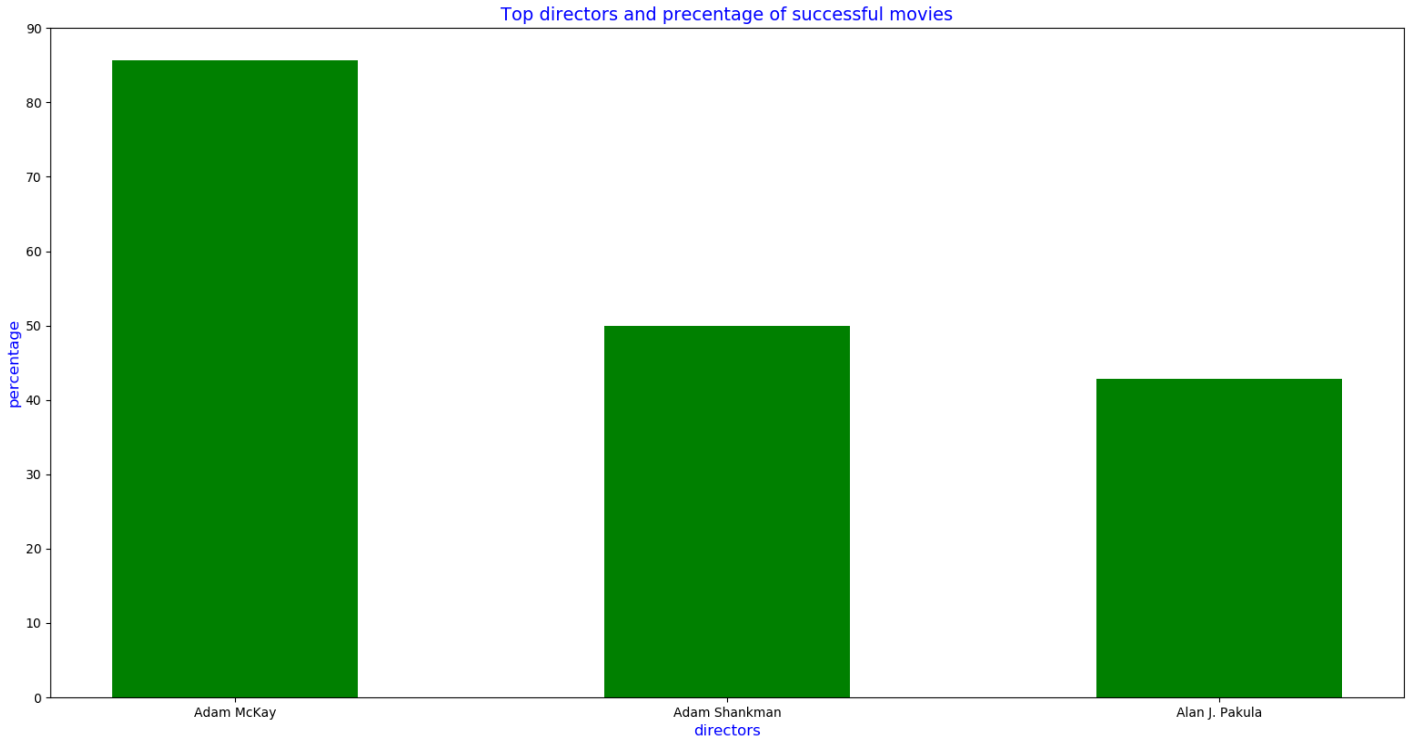
Line plot

Scatter plot

Summary statistics and plots :

Plots :





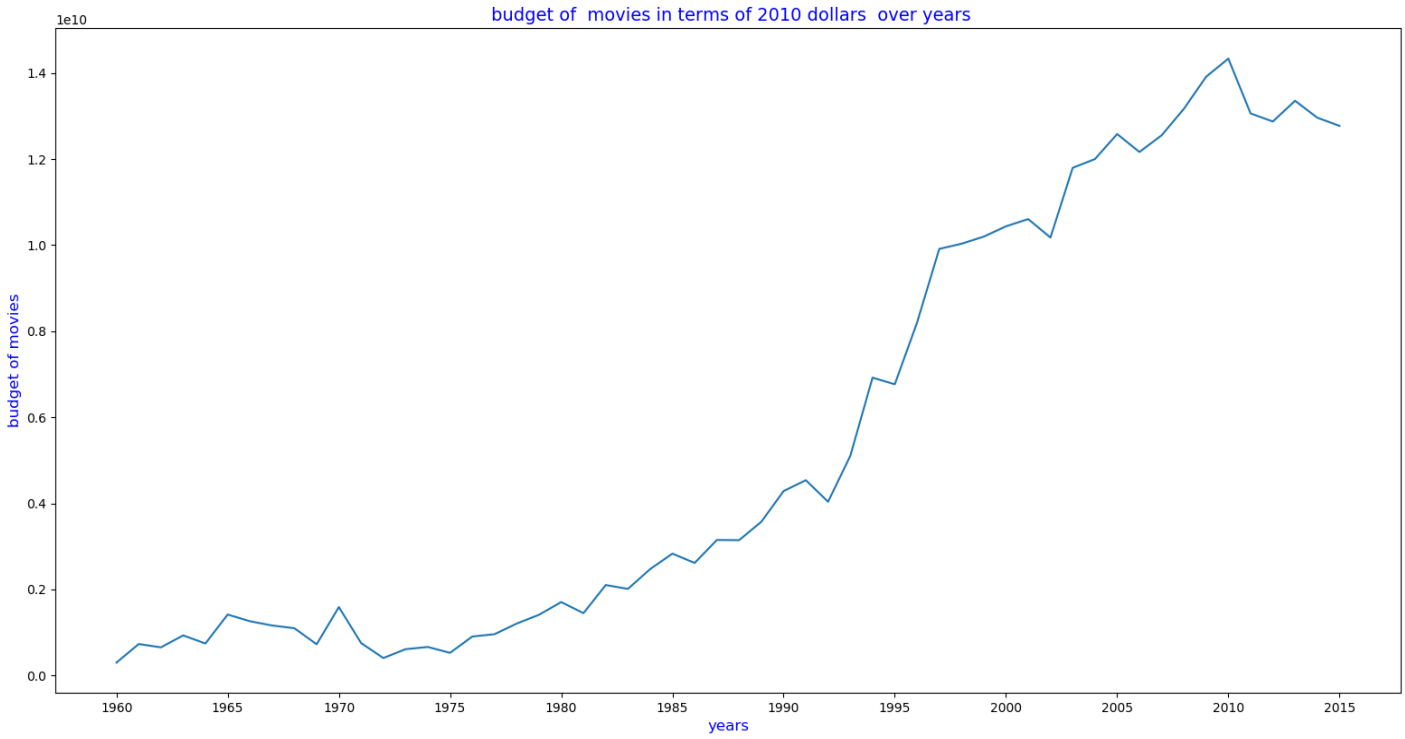
director	number of movies
Adam McKay	7
Adam Shankman	8
Alan J. Pakula	7

Adam Shankman made 6 successful movies with success 85 %

Adam Mckay made 6 successful movies with success 50 %

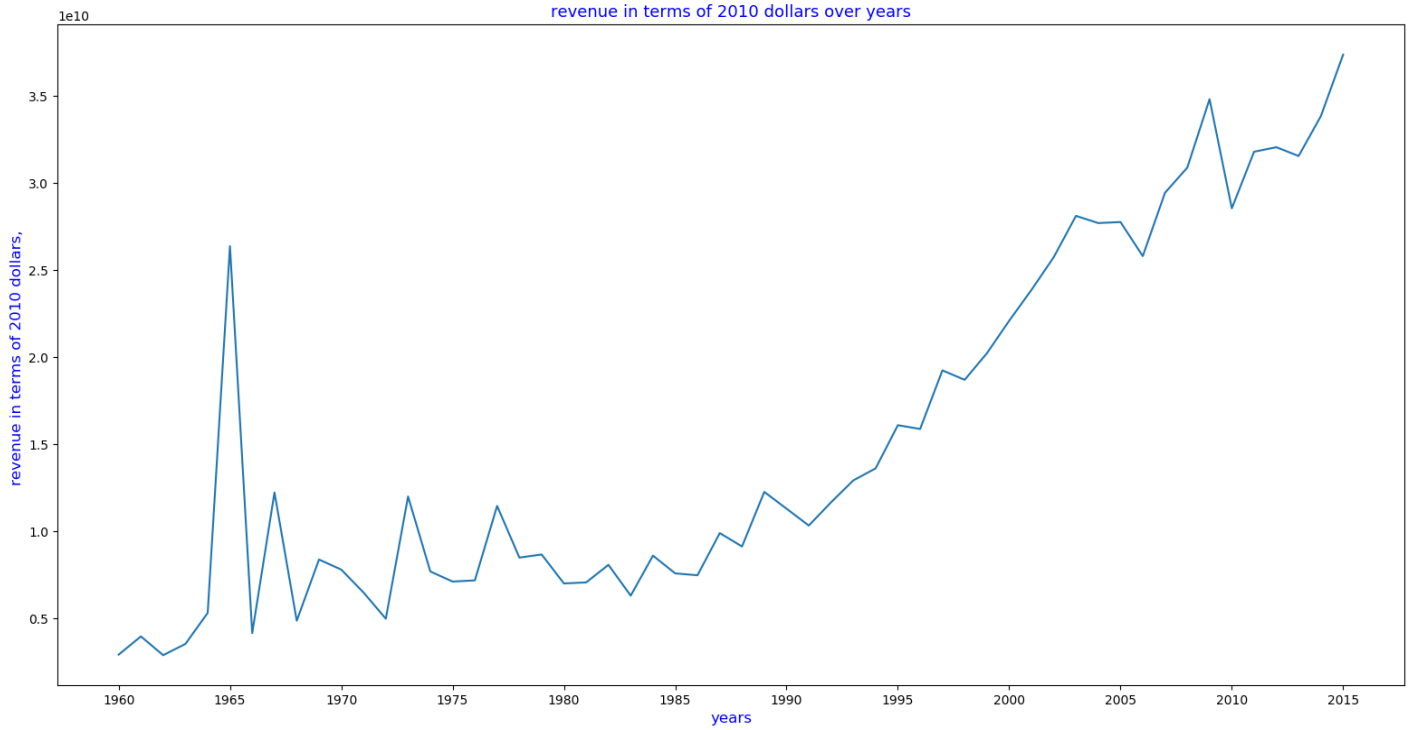
Alan J. Pakula made 3 successful movies with success 42.85 %

Adam Shankman is much better than Adam Mckay and Alan J. Pakula



Budget of movies increases over years.

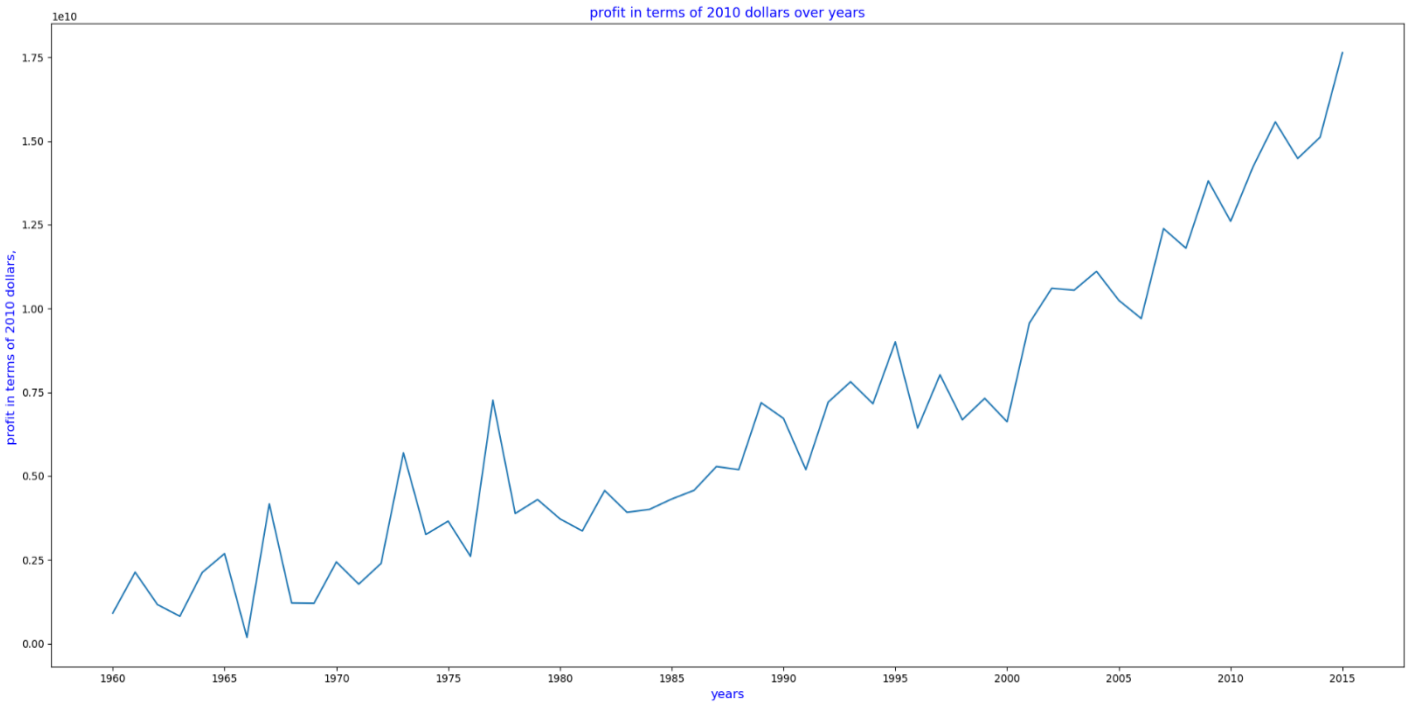
To make a good movie using new technology methods , budget of movies has increased.



Revenue of movies increases over years.

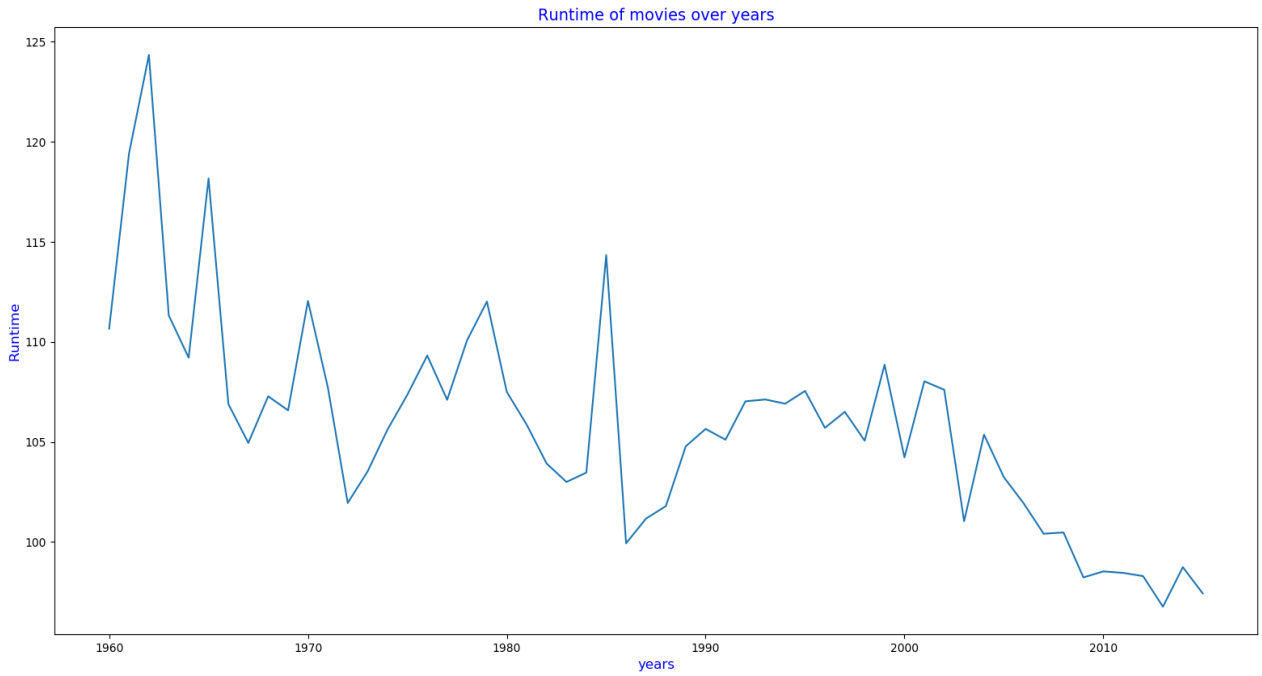
There has been more cinemas all over the world and more people go to cinemas.

Tickets price has increased .



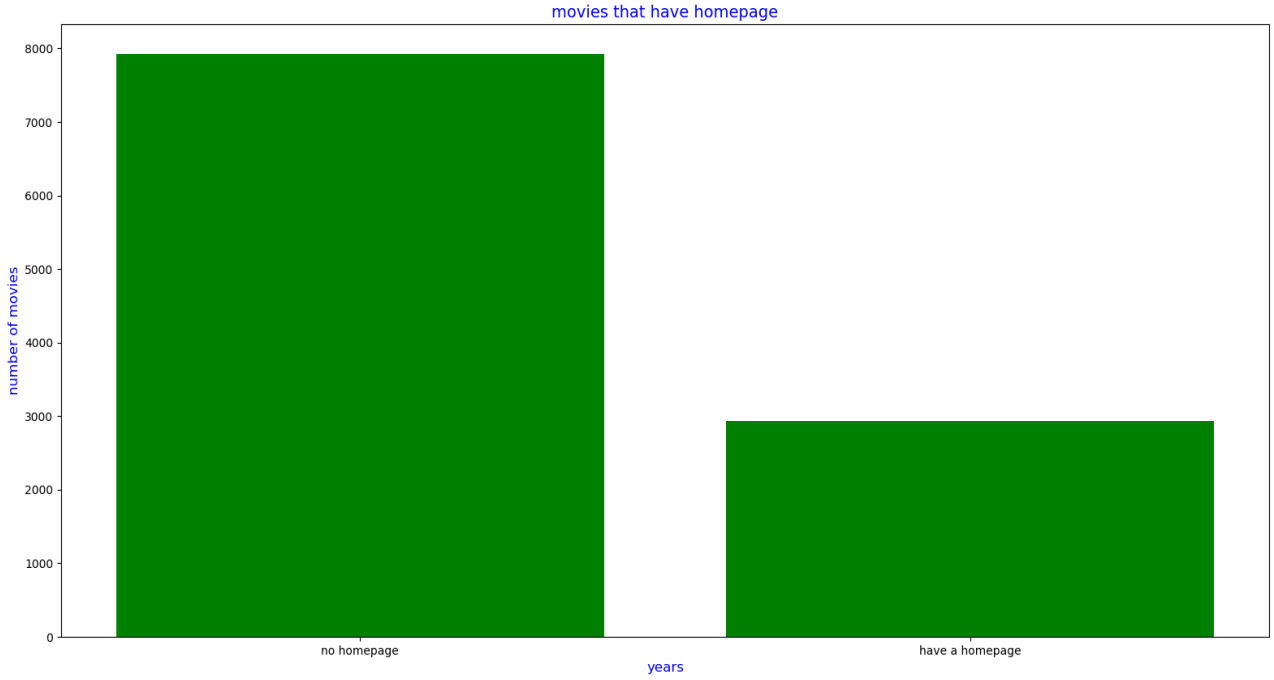
profit of movies increases over years.

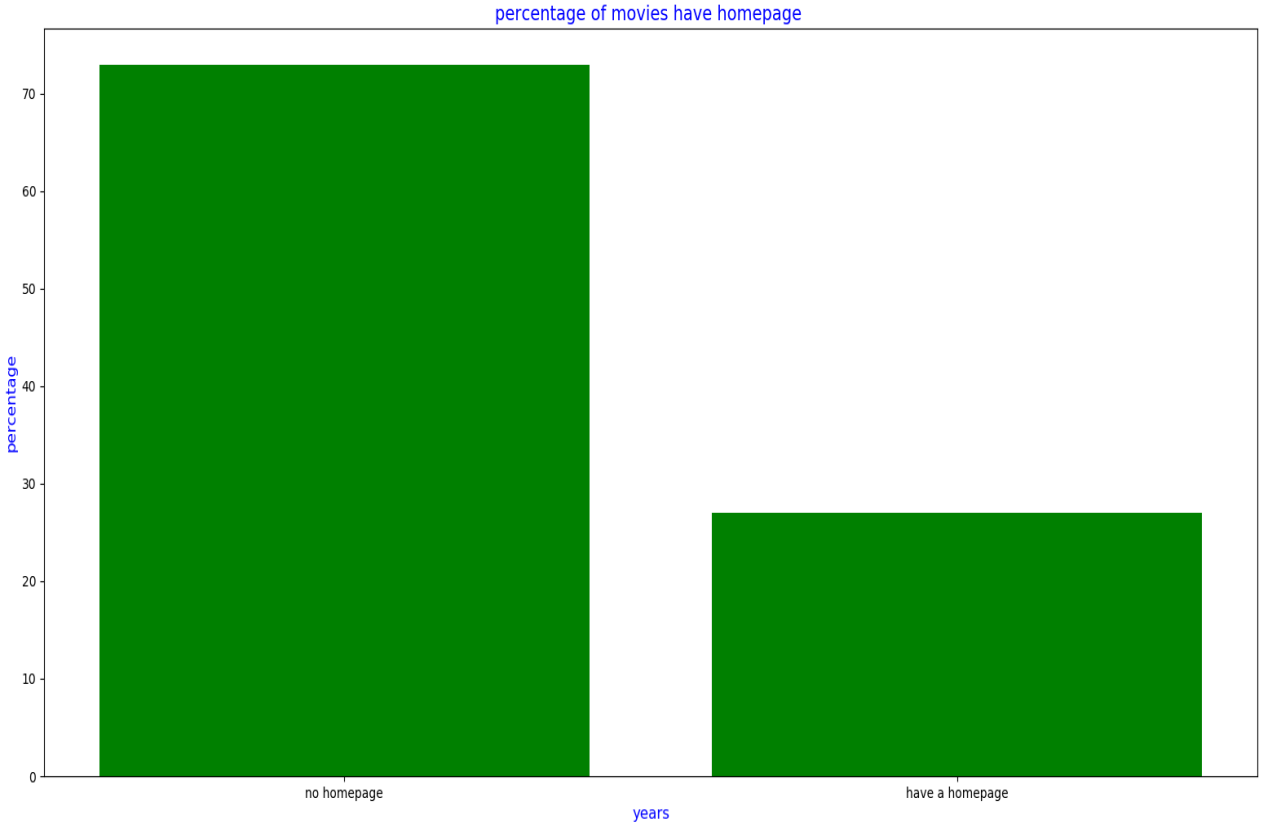
There has been more cinemas all over the world and more people go to cinemas.



Runtime of movies decreases over years.

Runtime of movies decreases to reduce the cost of the movie .Avoiding people get bored when the movie is long


movies that have homepage
number of movies
8000
7000
6000
5000
4000
3000
2000
1000
0
no homepage
have a homepage
years


percentage of movies have homepage
percentage
70
60
50
40
30
20
10
0
no homepage
have a homepage
years

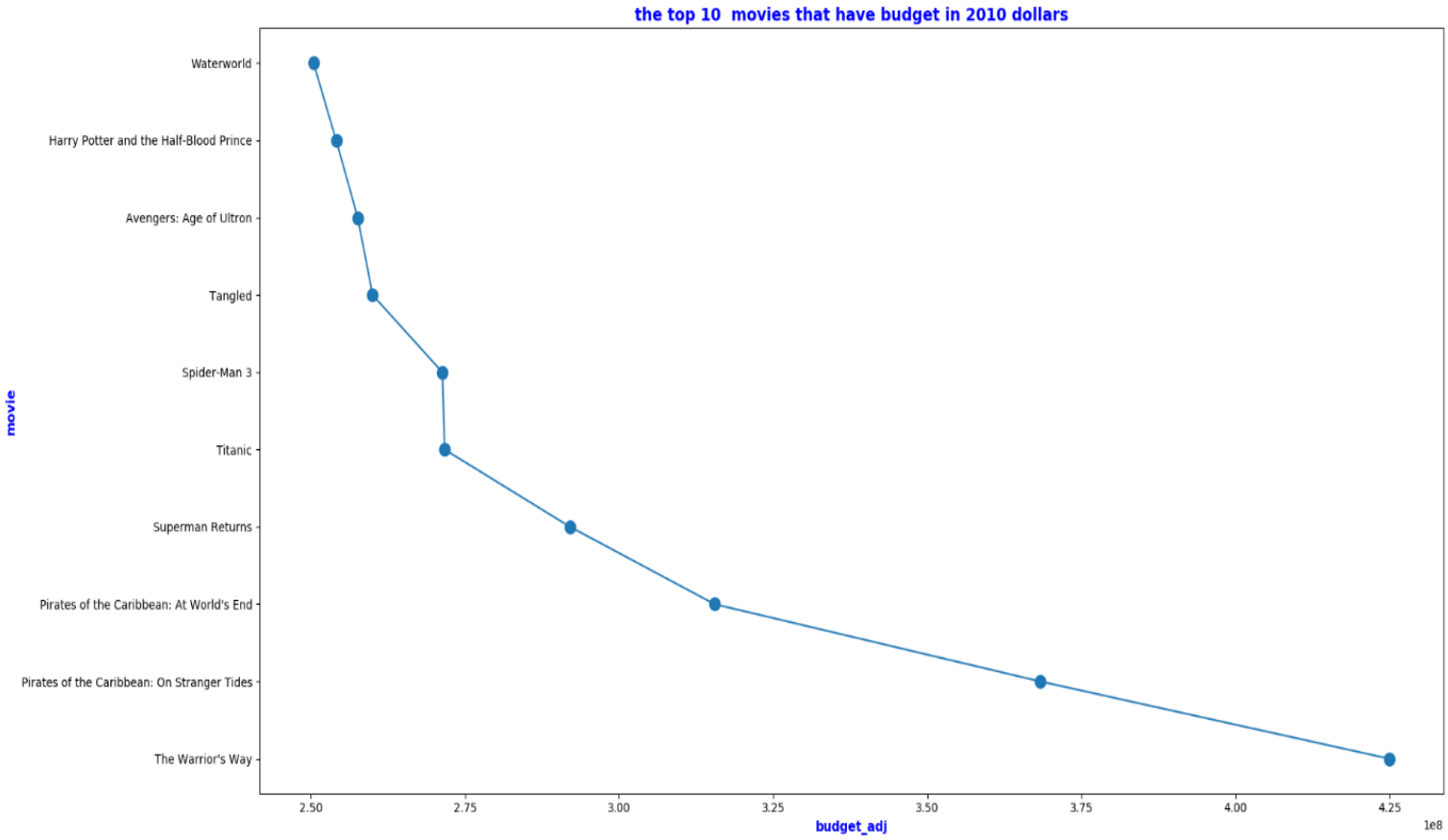
2939 movies have homepage with 27% of total number of movies

41.5 % of movies that have homepage are popular .

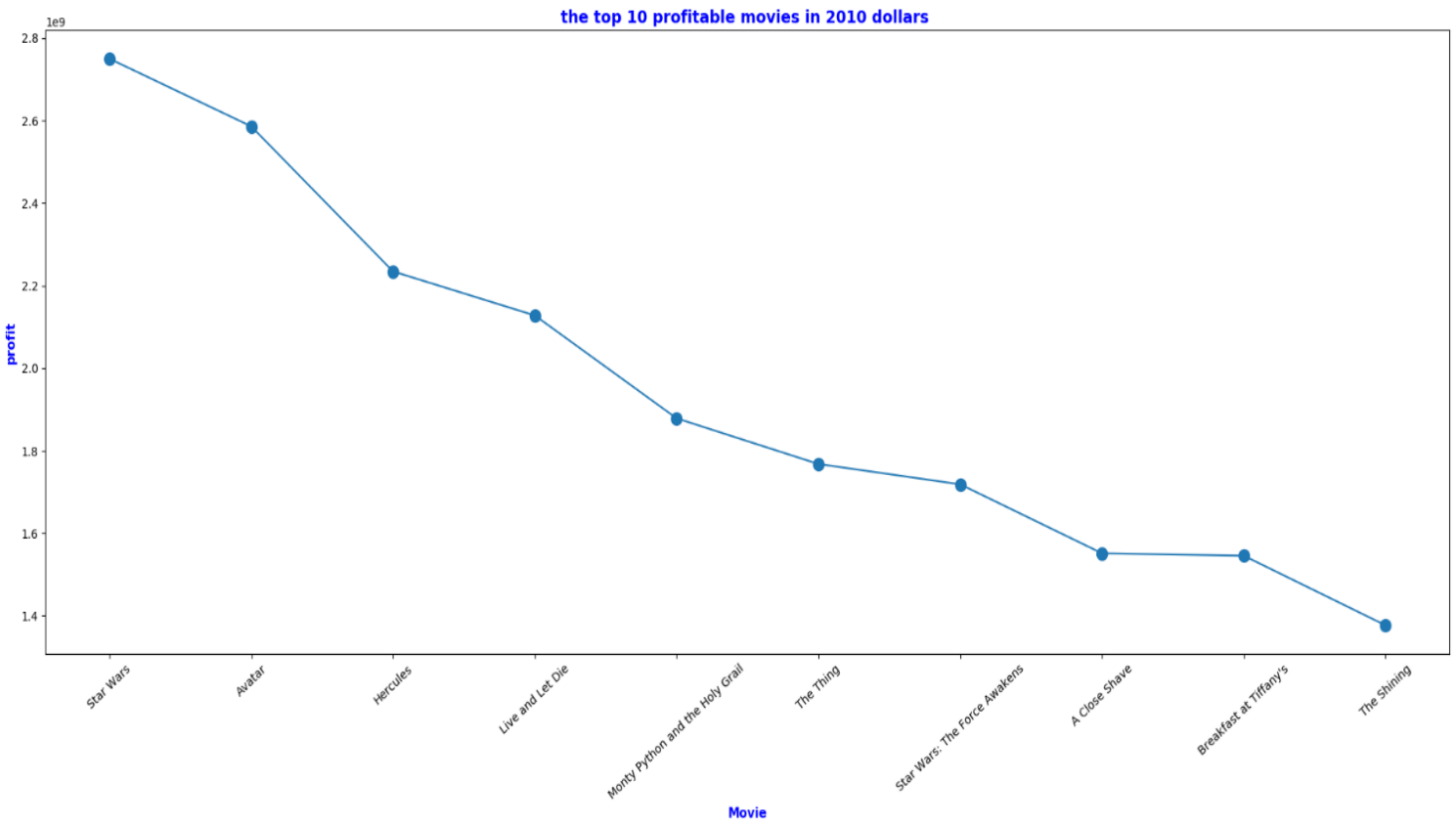
7929 movies don't have homepage with 73% of total number of movies

23.3 % of movies that don't have homepage are popular .

Having a homepage increase the propapbilty of being a popular movie.

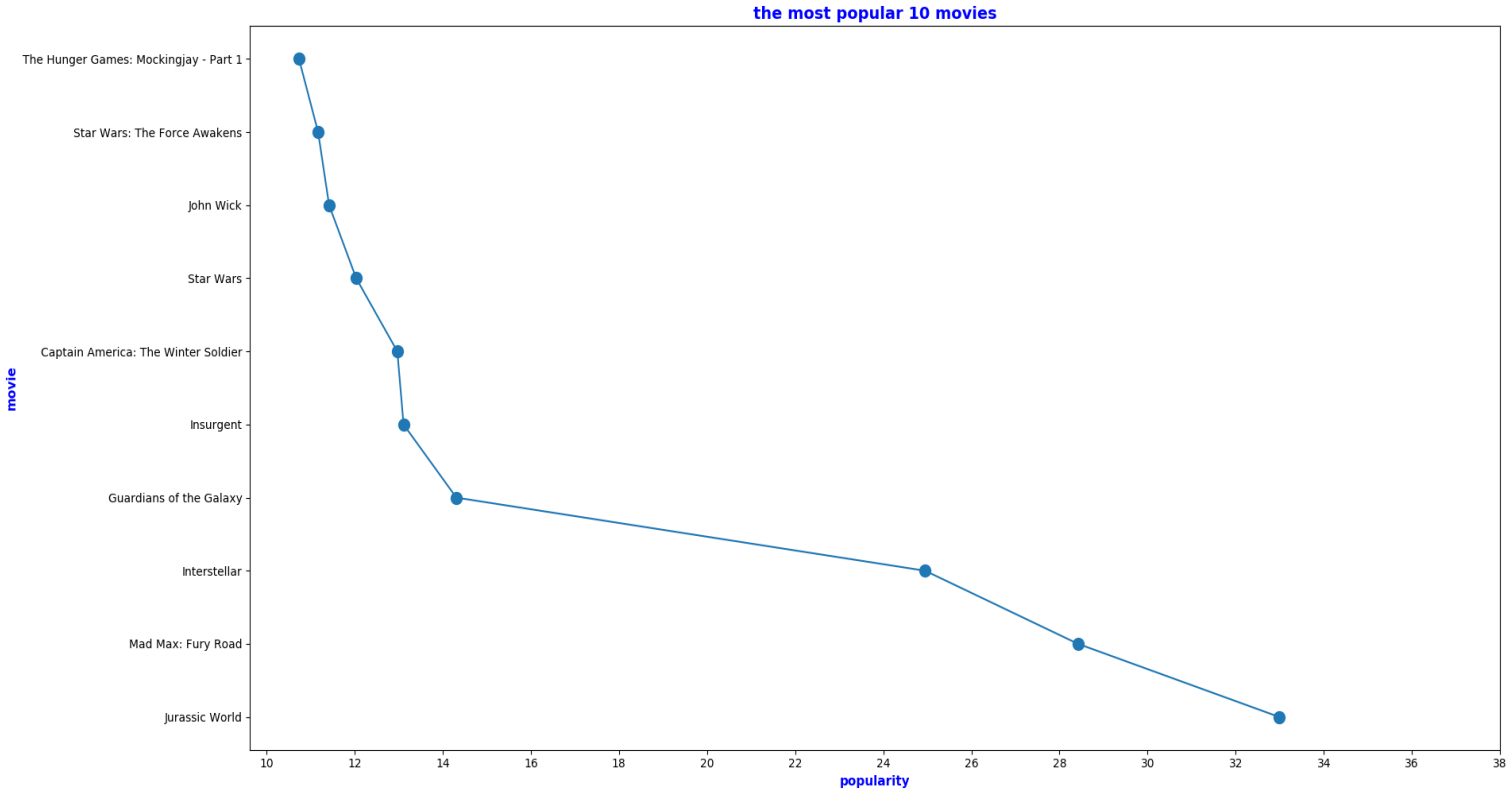


original_title	budget_adj
The Warrior's Way	4.250000e+0
Pirates of the Caribbean: On Stranger Tides	3.683713e+08
Pirates of the Caribbean: At World's End	3.155006e+08
Superman Returns	2.920507e+08
Titanic	2.716921e+08
Spider-Man 3	2.713305e+08
Tangled	2.600000e+08
Avengers: Age of Ultron	2.575999e+08
Harry Potter and the Half-Blood Prince	2.541001e+08
Waterworld	2.504192e+08



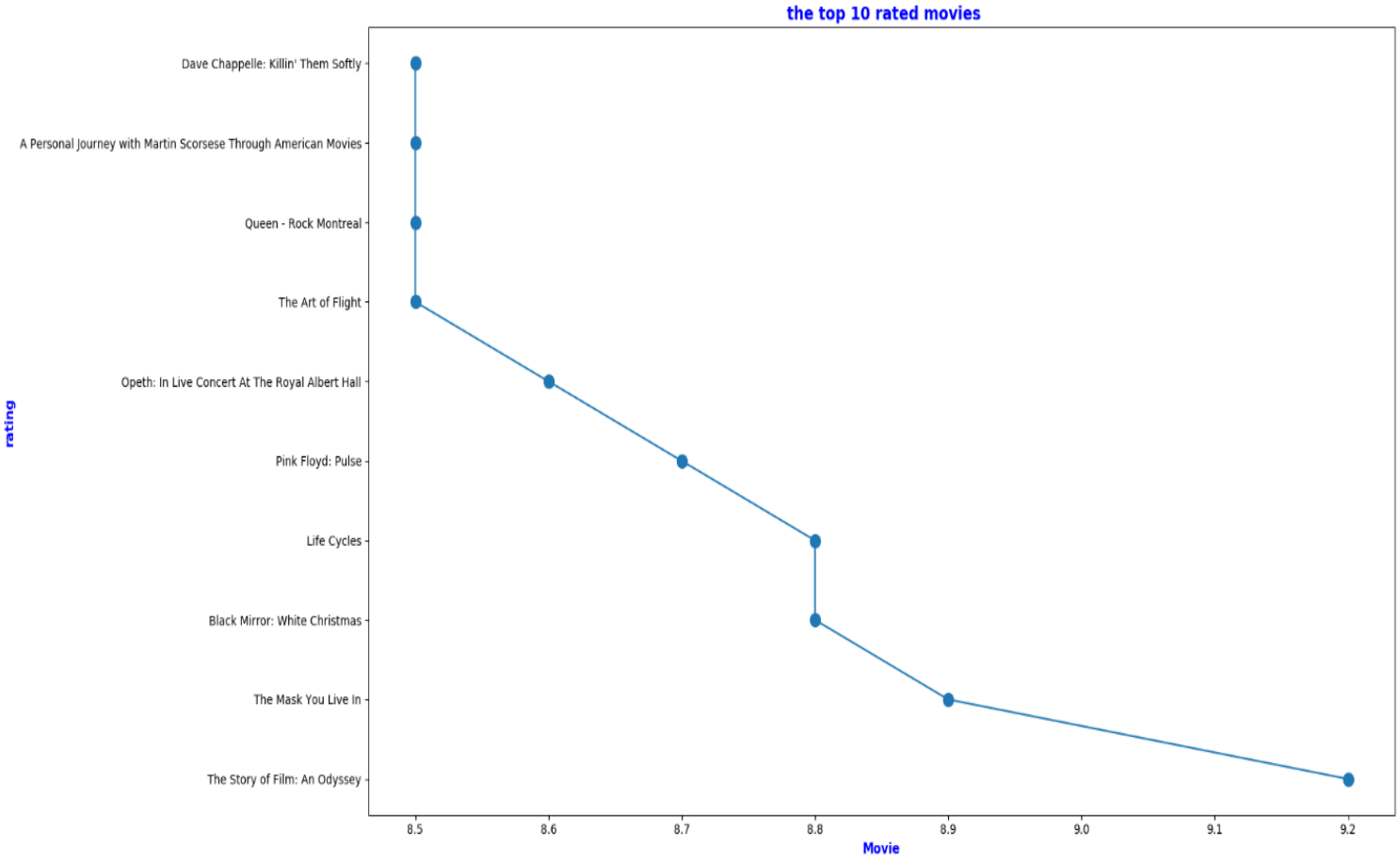
the most 10 profitable movies:

profit_adj	original_title
2.750137e+09	Star Wars
2.586237e+09	Avatar
2.234714e+09	Hercules
2.128036e+09	Live and Let Die
1.878643e+09	Monty Python and the Holy Grail
1.767968e+09	The Thing
1.718723e+09	Star Wars: The Force Awakens
1.551568e+09	A Close Shave
1.545635e+09	Breakfast at Tiffany's
1.376998e+09	The Shining



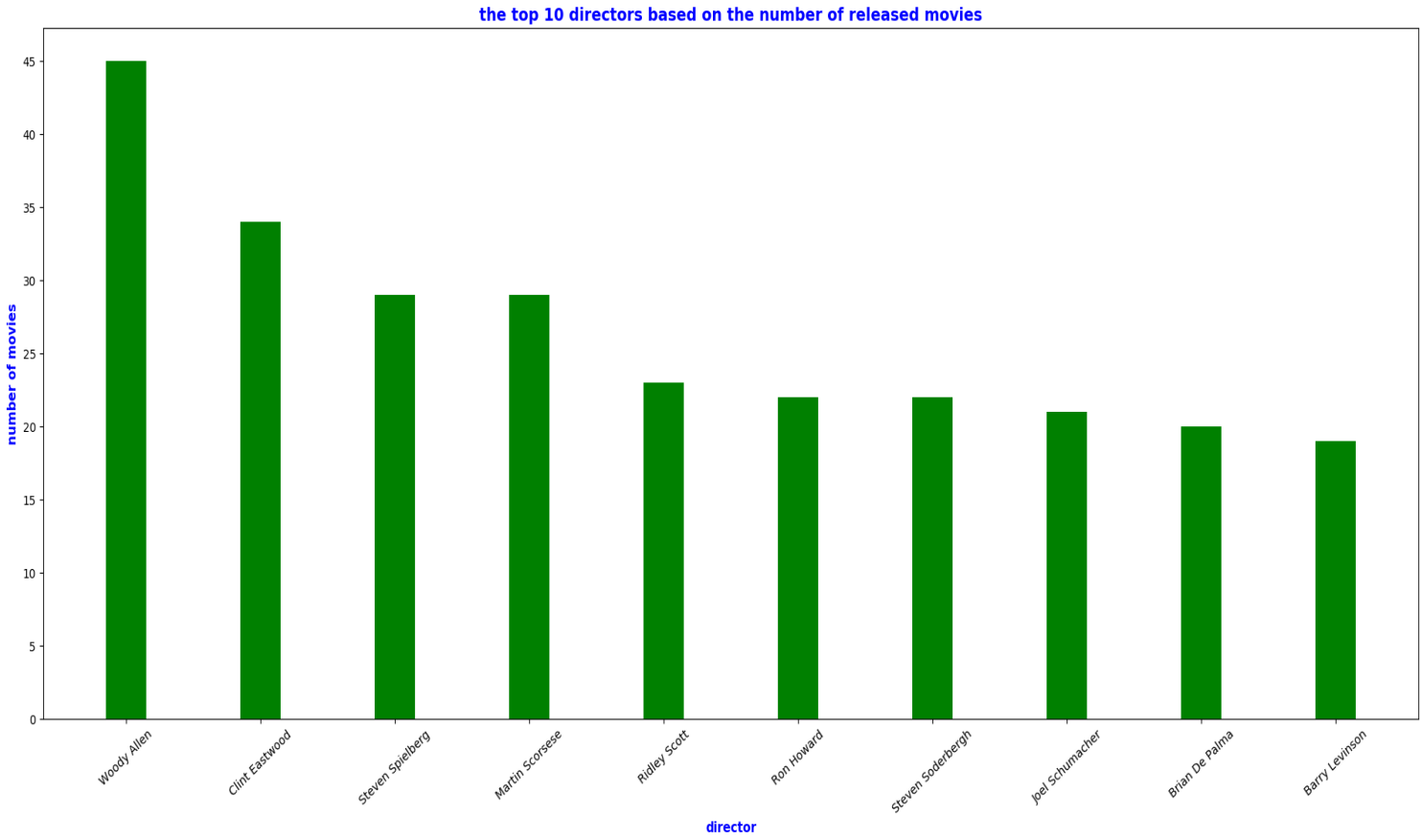
the most 10 popular movies:

popularity	original_title
32.985763	Jurassic World
28.419936	Mad Max: Fury Road
24.949134	Interstellar
14.311205	Guardians of the Galaxy
13.112507	Insurgent
12.971027	Captain America: The Winter Soldier
12.037933	Star Wars
11.422751	John Wick
11.173104	Star Wars: The Force Awakens
10.739009	The Hunger Games: Mockingjay - Part 1



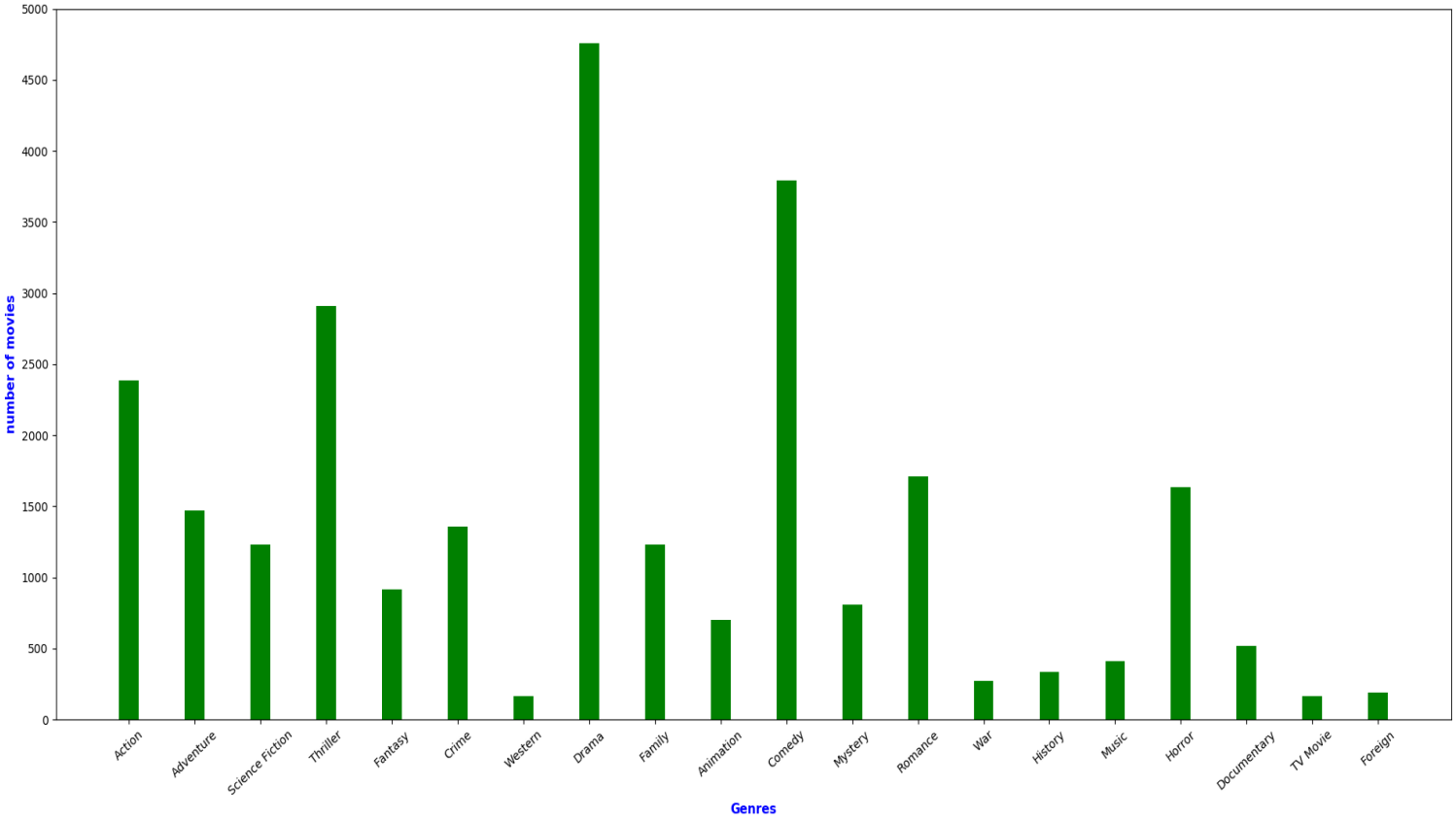
the top 10 rated movies

vote_average	original_title
9.2	The Story of Film: An Odyssey
8.9	The Mask You Live In
8.8	Black Mirror: White Christmas
8.8	Life Cycles
8.7	Pink Floyd: Pulse
8.6	Opeth: In Live Concert At The Royal Albert Hall
8.5	The Art of Flight
8.5	Queen - Rock Montreal
8.5	A Personal Journey with Martin Scorsese Throug...
8.5	Dave Chappelle: Killin' Them Softly



the top 10 directors based on the number of released movies

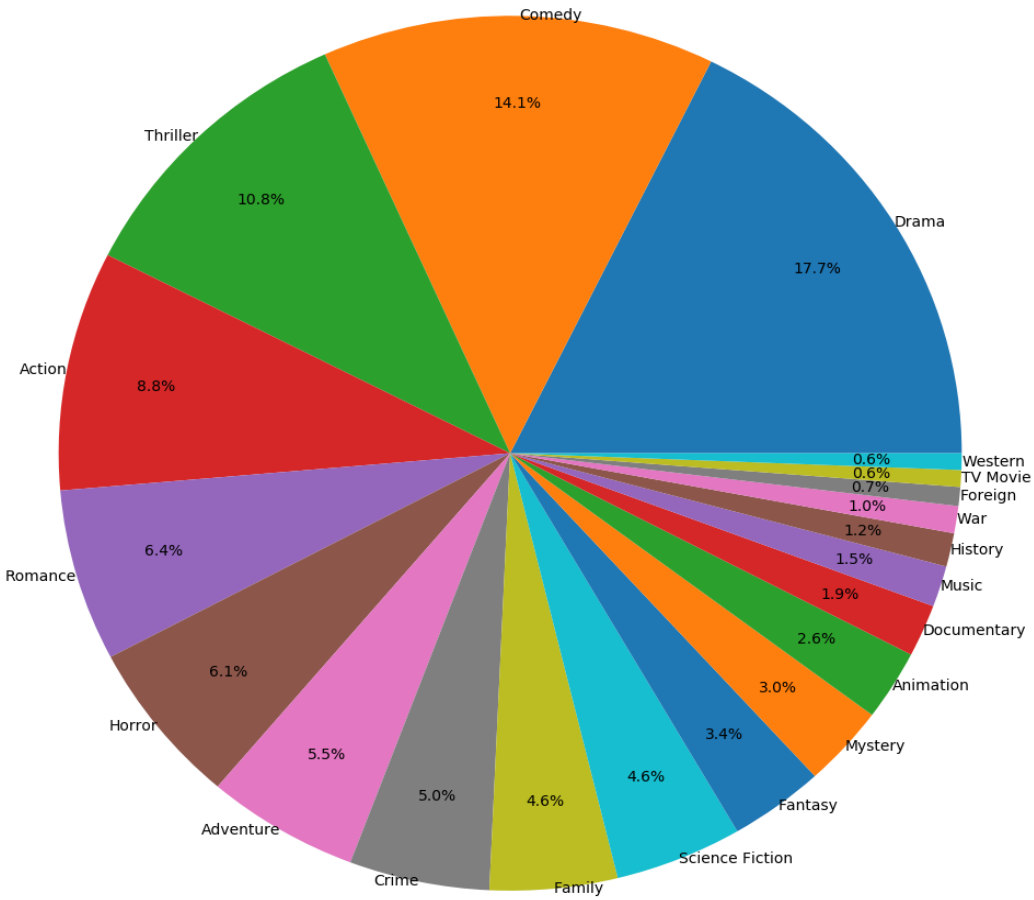
director	number of movies
Woody Allen	45
Clint Eastwood	34
Steven Spielberg	29
Martin Scorsese	29
Ridley Scott	23
Steven Soderbergh	22
Ron Howard	22
Joel Schumacher	21
Brian De Palma	20
Wes Craven	19



number of movies in each genre :

genre	number of movies
Drama	4760
Comedy	3793
Thriller	2907
Action	2384
Romance	1712
Horror	1637
Adventure	1471
Crime	1354
Family	1231
Science Fiction	1229
Fantasy	916
Mystery	810
Animation	699
Documentary	520

Music	408
History	334
War	270
Foreign	188
TV Movie	167
Western	165



movies in each genre

From 1960 to 2015 the top 5 genres based on the number of released movies :

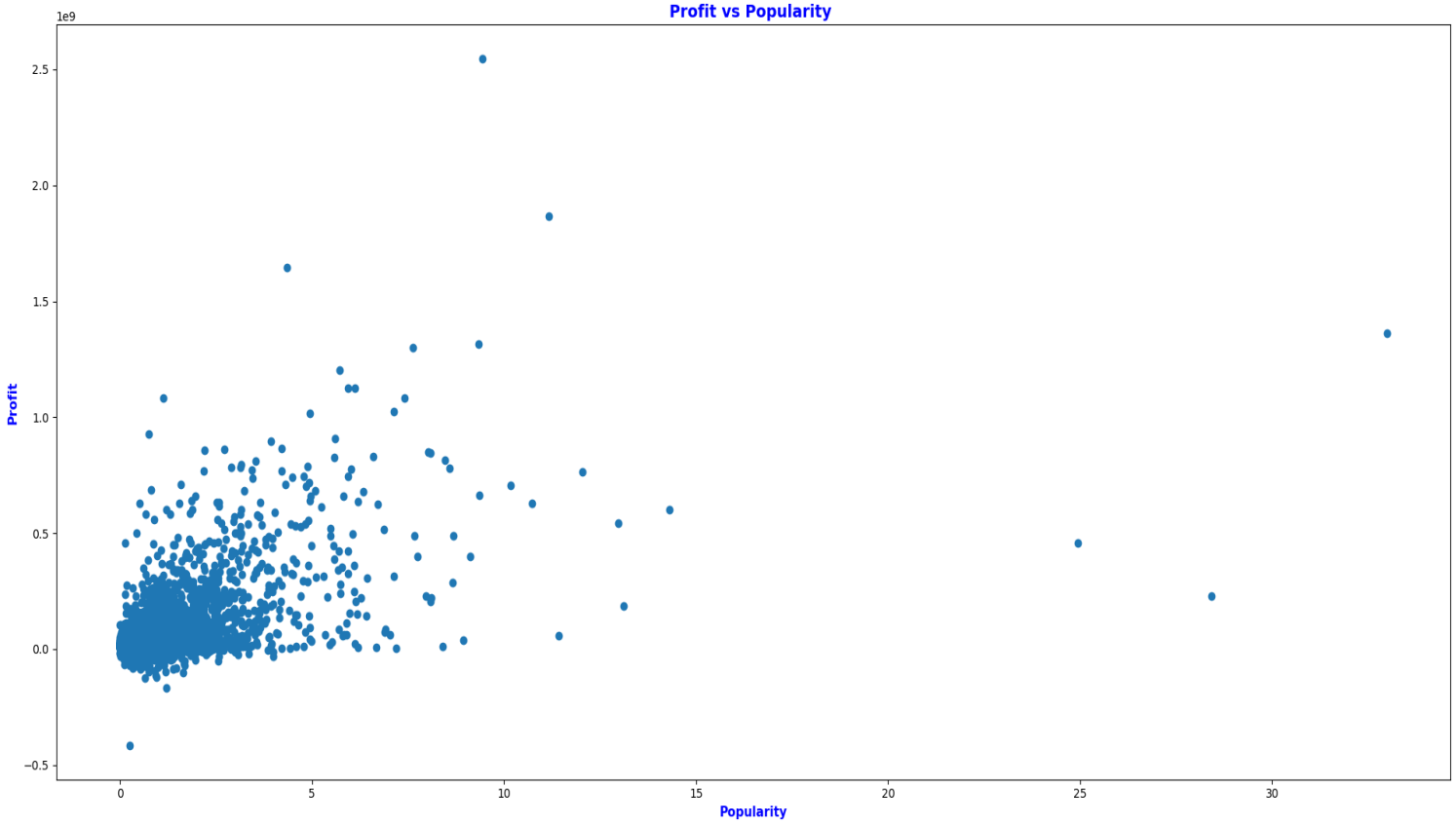
Drama with 4760 movies

Comedy with 3793 movies

Thriller with 2907 movies

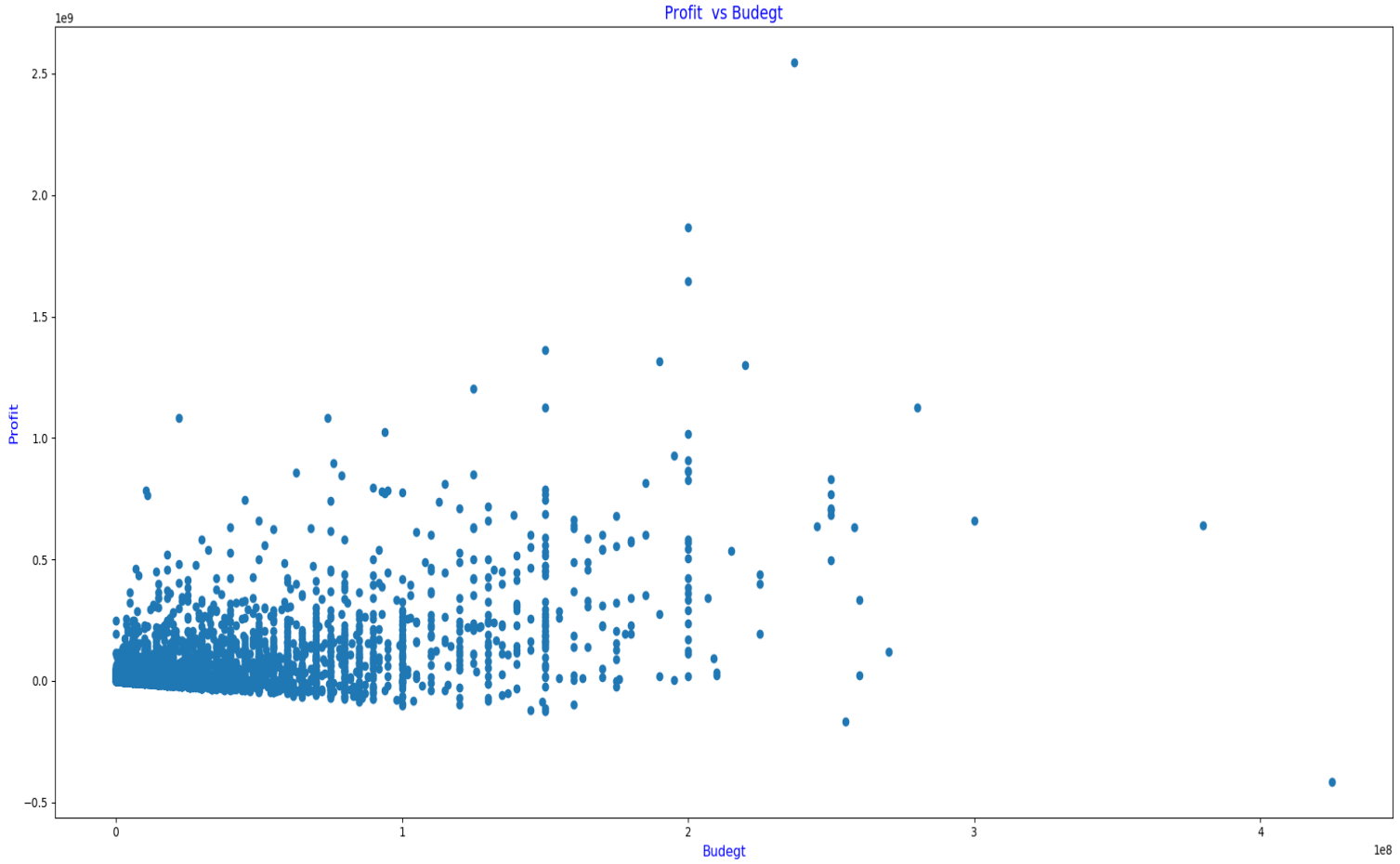
Action with 2384 movies

Romance with 1712 movies



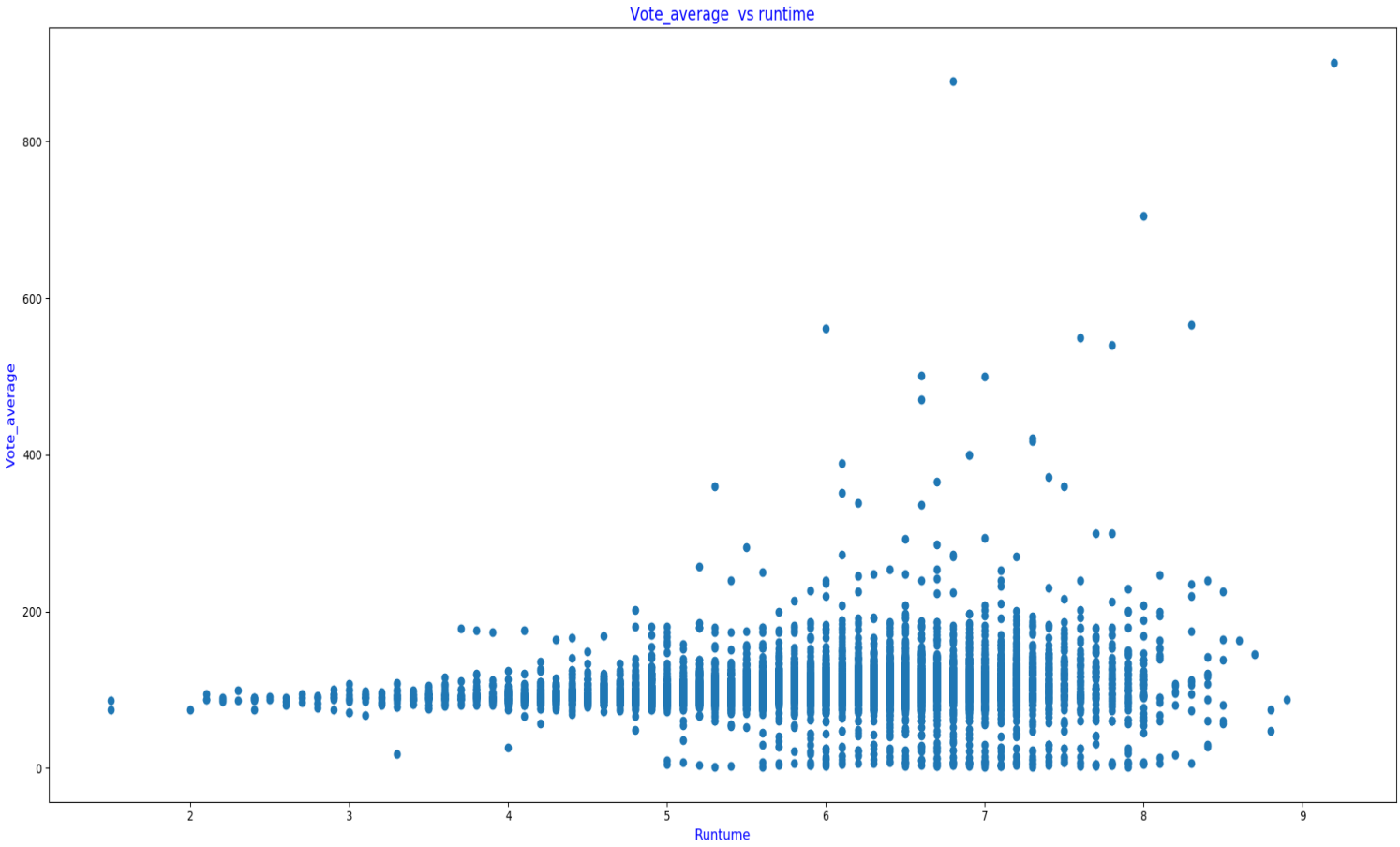
as popularity of movies increases , the profit of the movie increases

The Correlation coefficient between profit and popularity: 0.6098993604420108



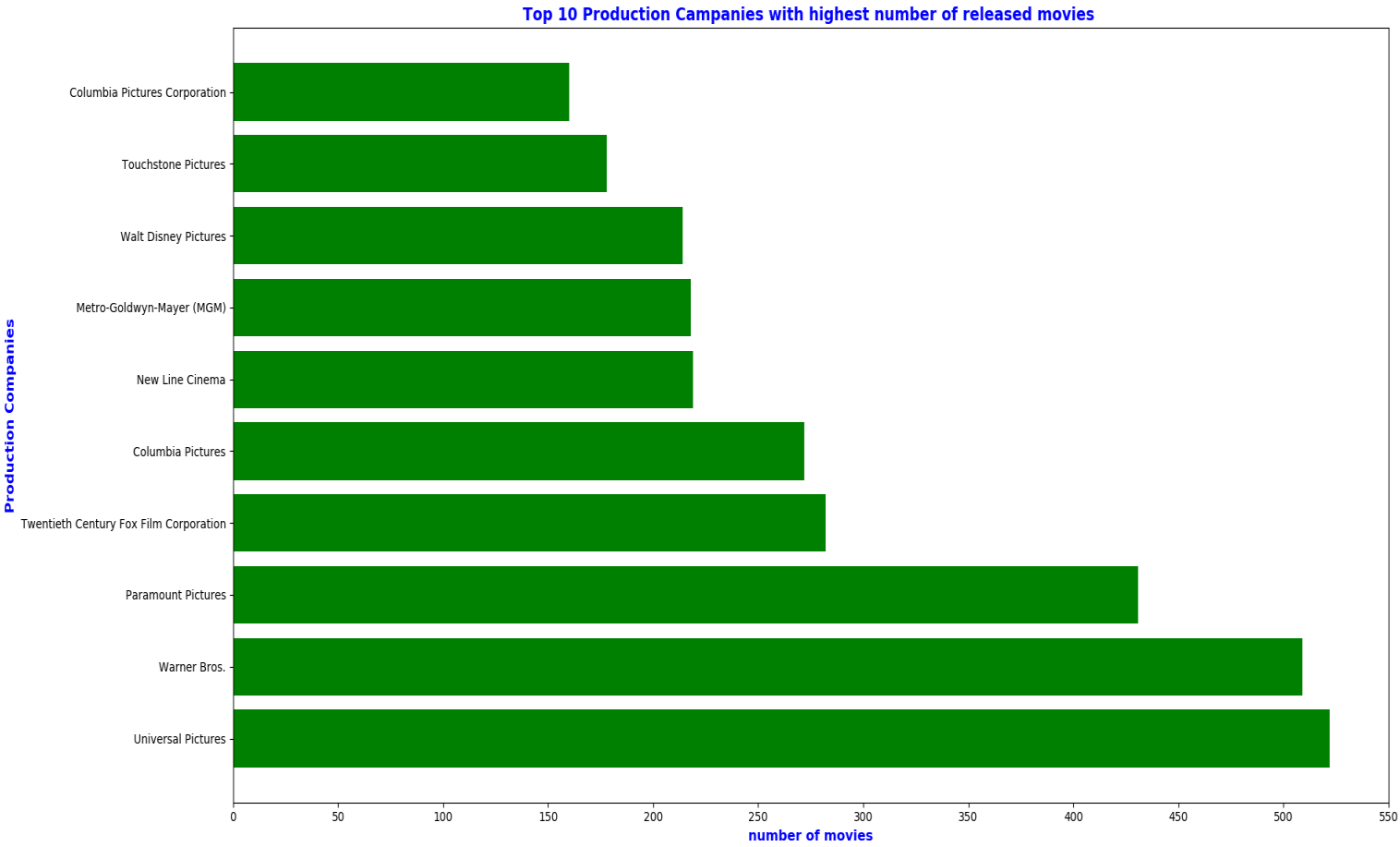
as budget of movies increases , the profit of the movie increase

The Correlation coefficient between Budegt and Profit: 0.5340505629877231



The Correlation coefficient between vote_average and runtume: 0.15485469873143615

There is no obvious strong relation between runtime and rating of the movie



top 10 production companies based on number of released movies:

Company	number of movies
Universal Pictures	522
Warner Bros.	509
Paramount Pictures	431
Twentieth Century Fox Film Corporation	282
Columbia Pictures	272
New Line Cinema	219
Metro-Goldwyn-Mayer (MGM)	218
Walt Disney Pictures	214
Touchstone Pictures	178
Columbia Pictures Corporation	160

conclusions :

1) The Warrior's Way : The movie that has the highest budget based on a reference dollar (2010) with budget 425 million dollars.

2) Fear Clinic : The moive that has the lowest budget based on a reference dollar (2010) with budget 0.921091 dollars.

3) The budget of movies increaases over years

4) 2010 is the year that has the highest budget for movies with 14.335 billion 2010 dollars

5) The revenue of movies increaases over years

6) 2015 is the year that has the highest revenue with 37.36 billion 2010 dollars

7) The profit of movies increaases over years

8) 1960 is the year that has the highest profit with 24.95 billion 2010 dollars

9) Number of movies increases over years

10) 2014 is the year of the highest number of movies with 700 movie

11) Runtime of movies decreases over years

12) 2013 is the year that has the lowest average of runtime of the movie with average runtime 96.7 minutes

13) Fresh Guacamole : is the shortest movie with 2 minutes

14) The Story of Film: An Odyssey : is the longest movie with 900 minutes

15) Having a homepage increase the propapbilty of being a popular movie but don't guarantee to be popular

16) popularity and profit are highly positivly correlated

17) budget and profit are positivly correlated

18) There is no effect of the runtime of movie and rating

19) Being popular or having high profit don't guarantee to be top rated

Limitation:

1) The dataset has a high number of missing data in important columns, which may affect the accuracy of the analysis results . For example the budget and revenue columns have many zeros which affect any analysist that includes financial itmes like profit , top printable movies , movies with the highest budget and so on