Data advanced project 1

TMDb movie data

This data set contains information about 10,000 movies collected from The Movie Database (TMDb), including user ratings and revenue.

● Certain columns, like 'cast' and 'genres', contain multiple values separated by pipe (|) characters. ● There are some odd characters in the 'cast' column. Don't worry about cleaning them. You can leave them as is.

● The final two columns ending with "_adj" show the budget and revenue of the associated movie in terms of 2010 dollars, accounting for inflation over time.

I want to show some correlations and insights

We have some question :

-who are the Most of the director made films ?

-What are the most and the least popular movies?

- What are the most and the least film's revenue_adj?

- Has the film industry developed?

- Does the budget have a big impact on popularity?

->Before answering these questions we should preprocess this data

-Data Wrangling

. gather

. assess

. clean

- I import data from file.csv('tmdb-movies.csv')

- assess data to identify any problems in data’s quality or structure,

- I found one duplicated data, some useless columns, a lot of zeros in budget and revenue

- I took a copy from the data

-drop duplicated values

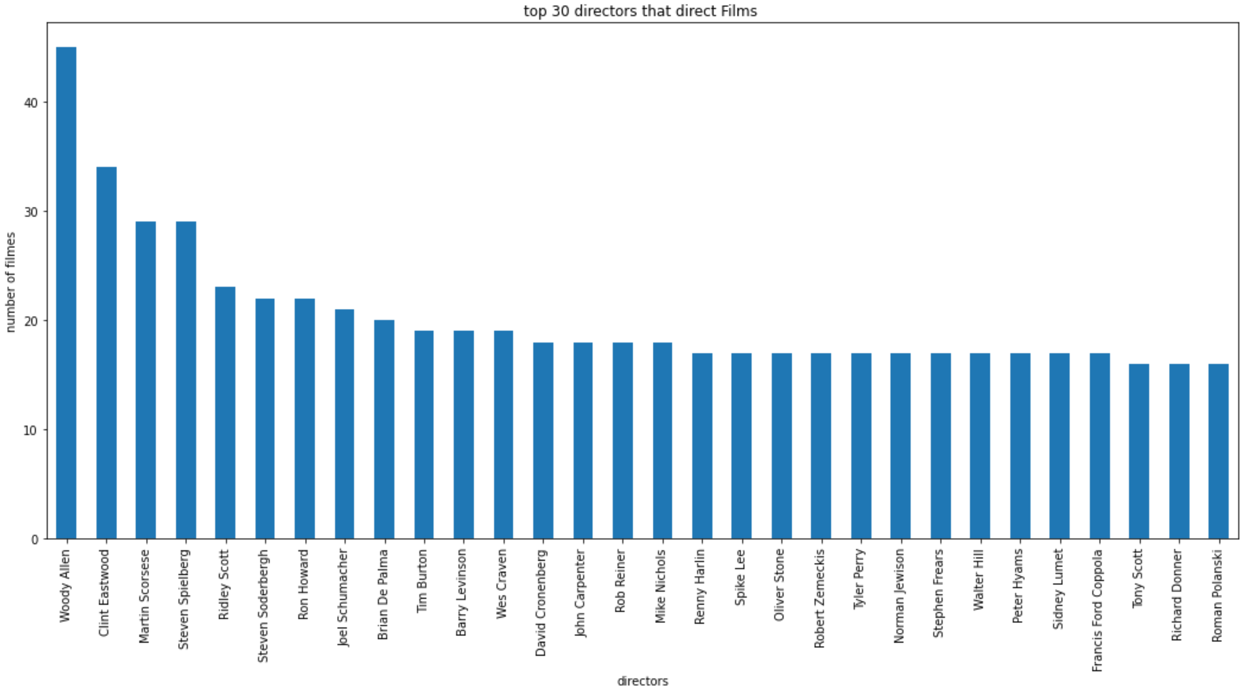
-remove useless columns

-replace zeros to Nan then dropped Nan

-who are the Most of the director made films ?

In this figure I show Top 30 directors that direct Films

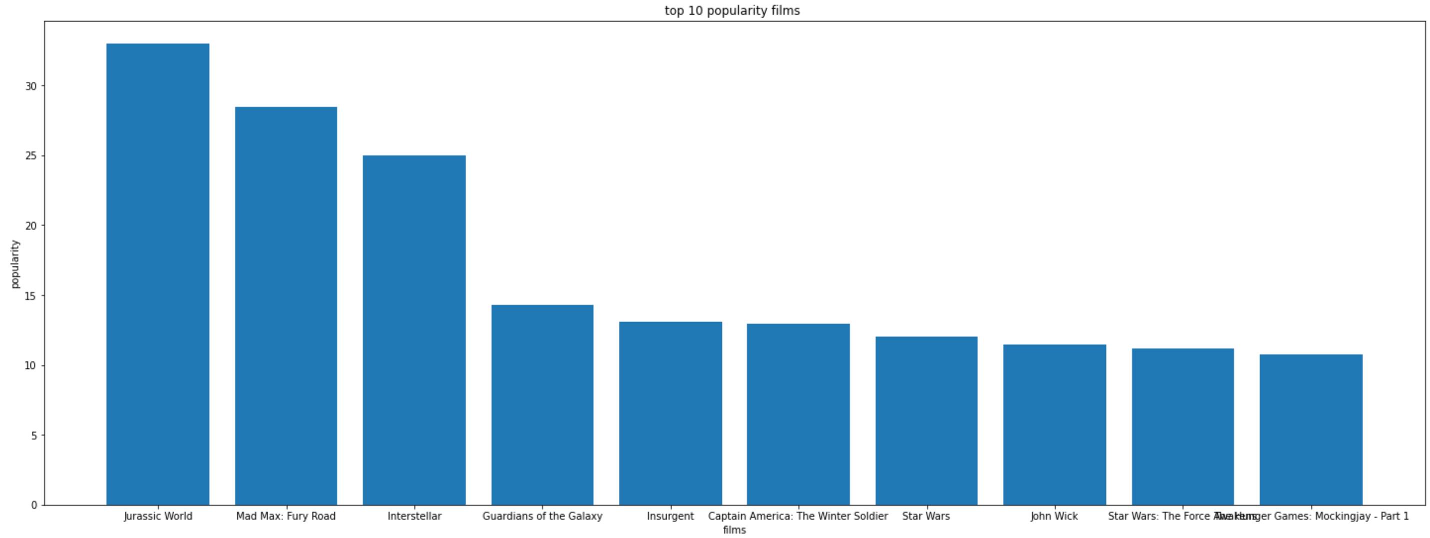
1- Woody Allen he made 45 films
2-Clint Eastwood he made 34 films
3-Martin Scorsese he made 29 films
Steven Spielberg he made 29 films



-This does not confirm that the 30 most directors who made films are the best 30 directors, But this does not prevent they have much experience in this field

-What are the most and the least popular movies?

I will show Top 10 popularity films



Most popular film: Jurassic World

(released 2015, 32.985763 popularity, 1513528810 revenue, to Colin Trevorrow director)

Least popular film: North and South, Book I

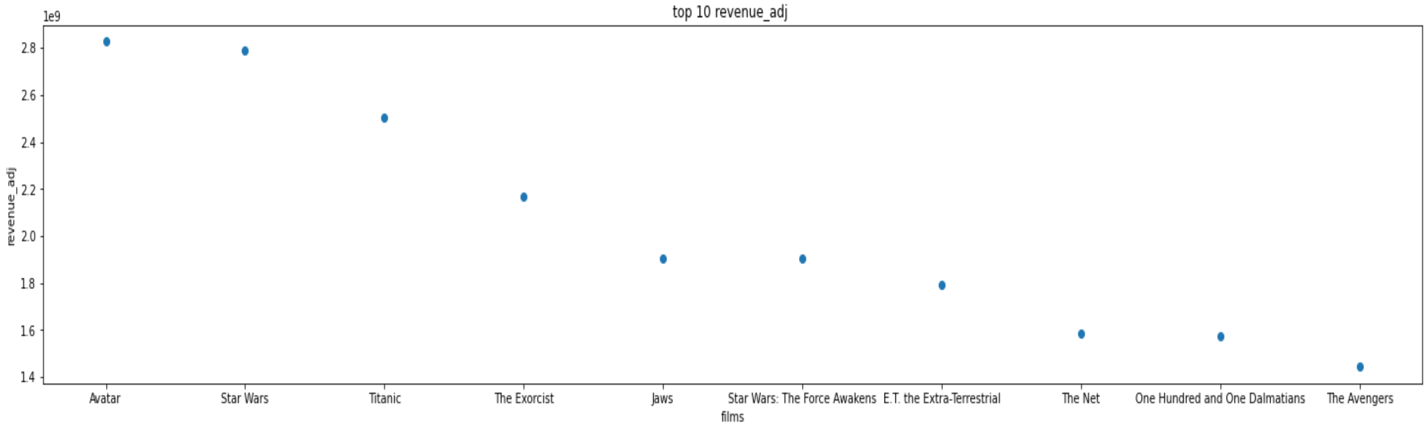
(released 1985, 0.000065 popularity)

The most popular should not be the highest revenue

But there is a direct relationship between the success of the movie’s popularity and the achievement of the highest revenues, and this is what we will discuss in the following sections

- What are the most and the least film's revenue_adj?

I will show top 10 revenue_adj



Film with the highest revenue_adj : avatar

(released 2009, 9.432768 popularity, 237000000 revenue, to James Cameron director)

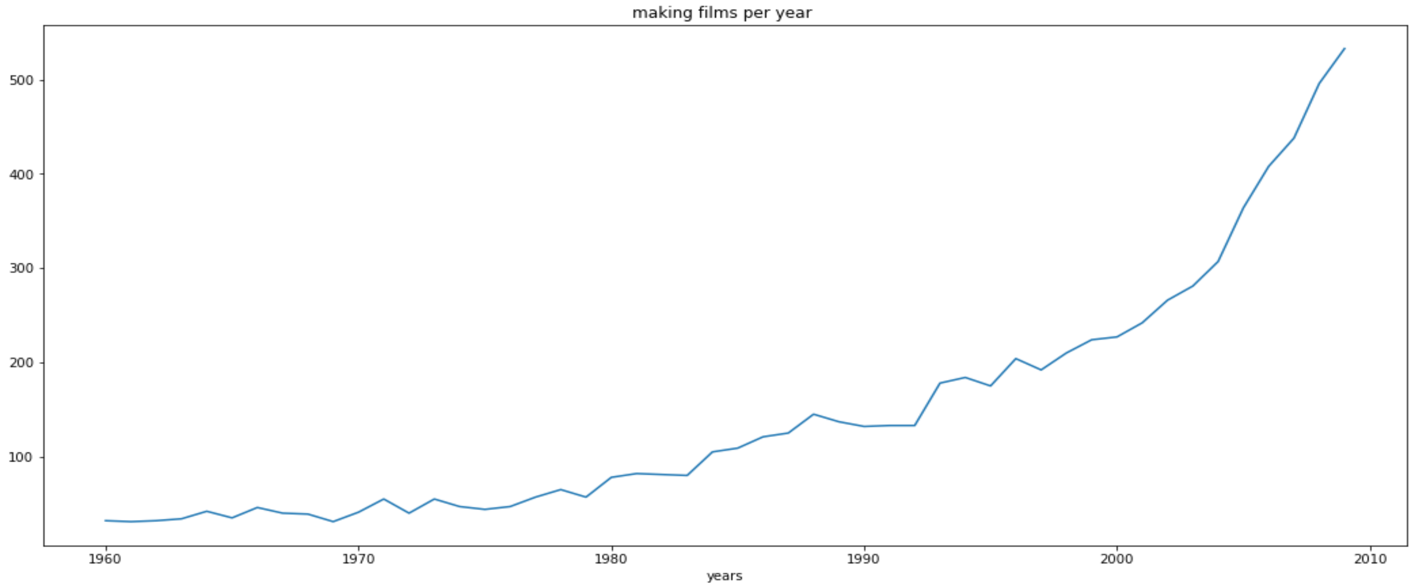
Film with lowest revenue_adj: Shattered Glass

(released 2003, 0.462609 popularity, 2.0 revenue, to Billy Ray director)

As we explained previously, the highest popularity does not have to be the highest revenue

the highest revenue film (avatar) does not have the highest popularity but have a good popularity

- Has the film industry developed?



direct correlation between released year and the number of films on its

Of course industry developed, because the number of films per year increases significantly, and we can see the best popularity in 5015,the highest revenue in 2009 .

This means increased competition in the film industry

- Does the budget have a big impact on popularity?

Has a medium direct correlation (pearson correlation coefficient)
= 0.4469075188988612

-> As we see lately the film industry is in a terrible development in terms of numbers and money.

Filmmakers and production companies should increase the budget because we noticed the direct relationship between movie popularity and budget.