(TMDB movie data)

Investigate a dataset

Overview

This data set contains movie information such as cast and genre, and also financial information regarding the movie's budget and revenue. It also includes adjusted budget and revenue values in terms of 2010 dollars, accounting for inflation over time. These columns are called 'budget_adj' and 'revenue_adj' respectively.

I answered some questions such as ...

- WHAT IS THE BEST MONTH TO RELEASE YOUR MOVIE?
- What is the best and the worst movies Best movies ?
- Who are the most popular director , actors , companies .
- COMMON COMBINATIONS BETWEEN ACTORS DIRECTORS AND COMPANIES
- MOST COMMON RELEASE MONTH FOR EVERY ACTOR ?
- Know common genre and determine the best
- What is genre distribution over years .

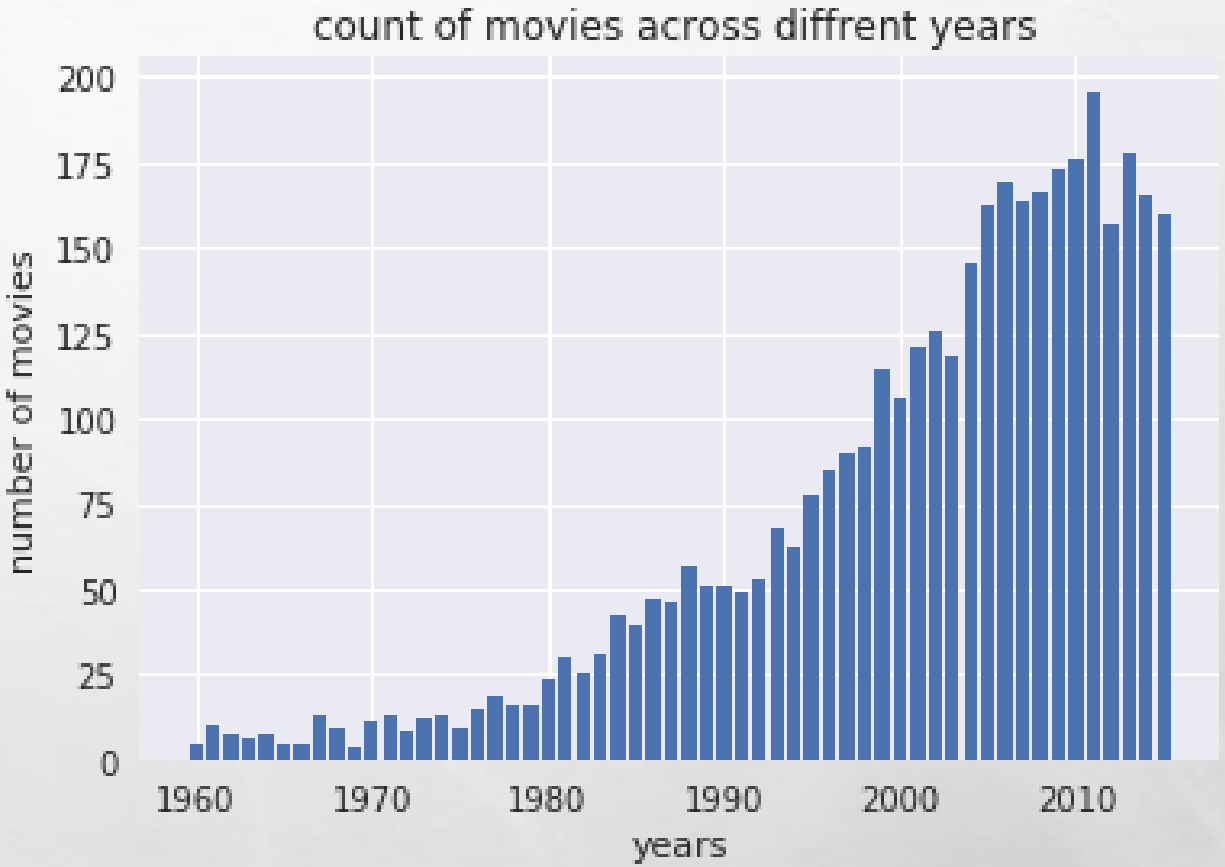
Number of movies per year distribution .

Its clearly that the distribution is left skewed that's mean the number of movie increase every year with exponential relation .

the year with the maximum number of movies is 2011

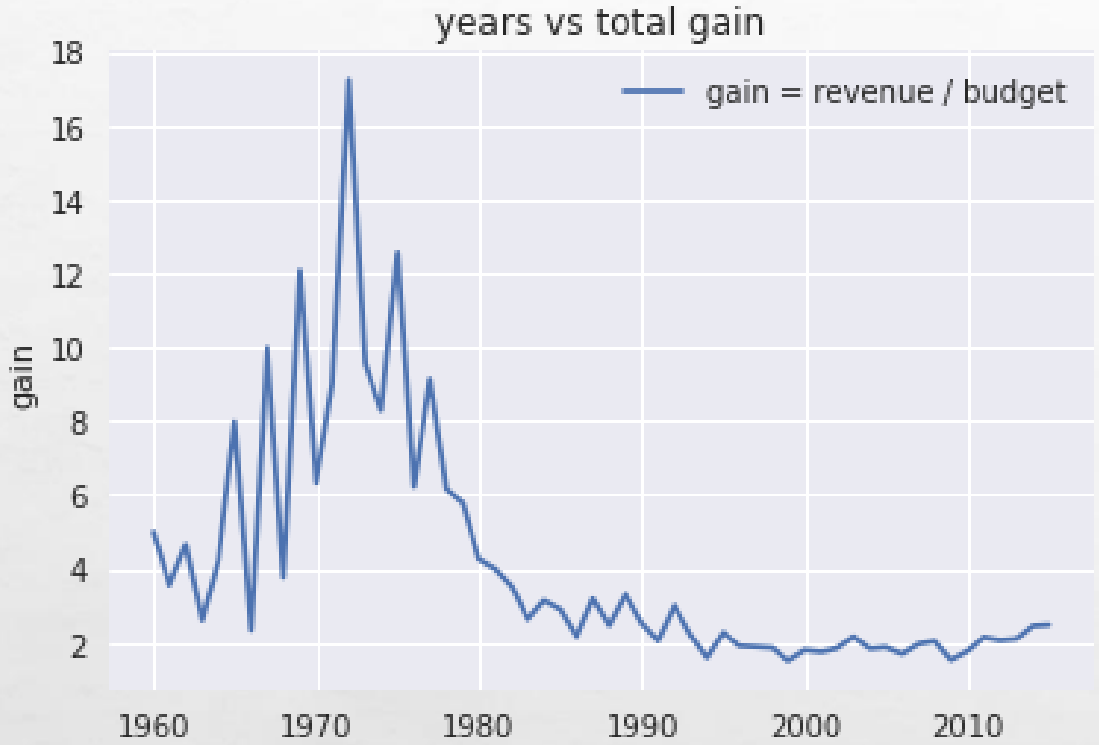
"2014 before cleaning"

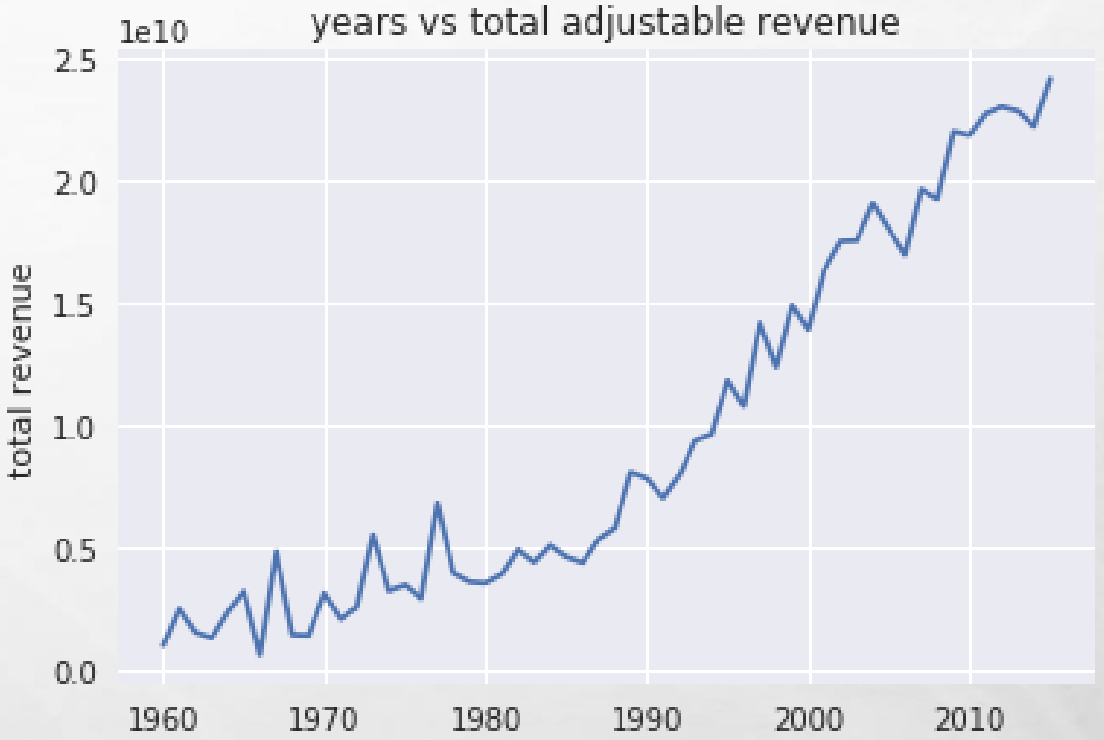
the year with the minimum number of movies is 1961



Note : this is not 100% accurate because some records are deleted at the cleaning process

WHAT WAS THE LUCKY YEAR ?





- This graph show that the heist gain was in 1972 that mean you could use small budget to gain high revenue but now to get more revenue you should but more budget .

- its clearly that the revenue has an exponential distribution and that make sense because many reasons such as marketing , income m peoples interests , number of movies …etc .

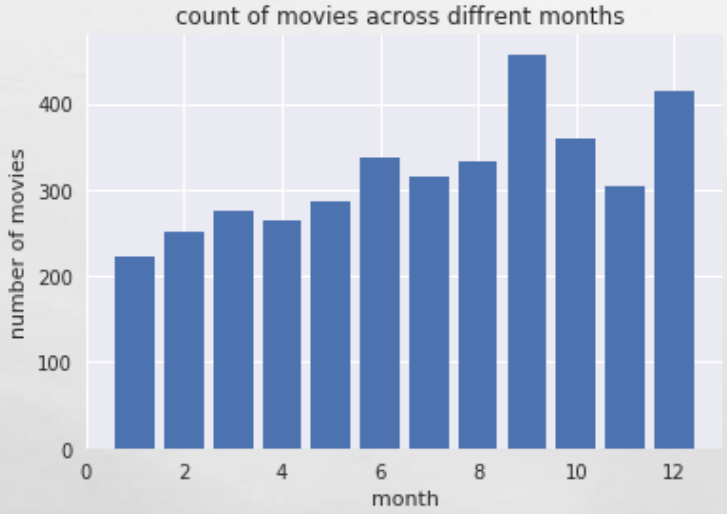
WHAT IS THE BEST MONTH ?

Obviously the best month to publish your movie is 6 or 12 because you have more chance to gain more revenue .

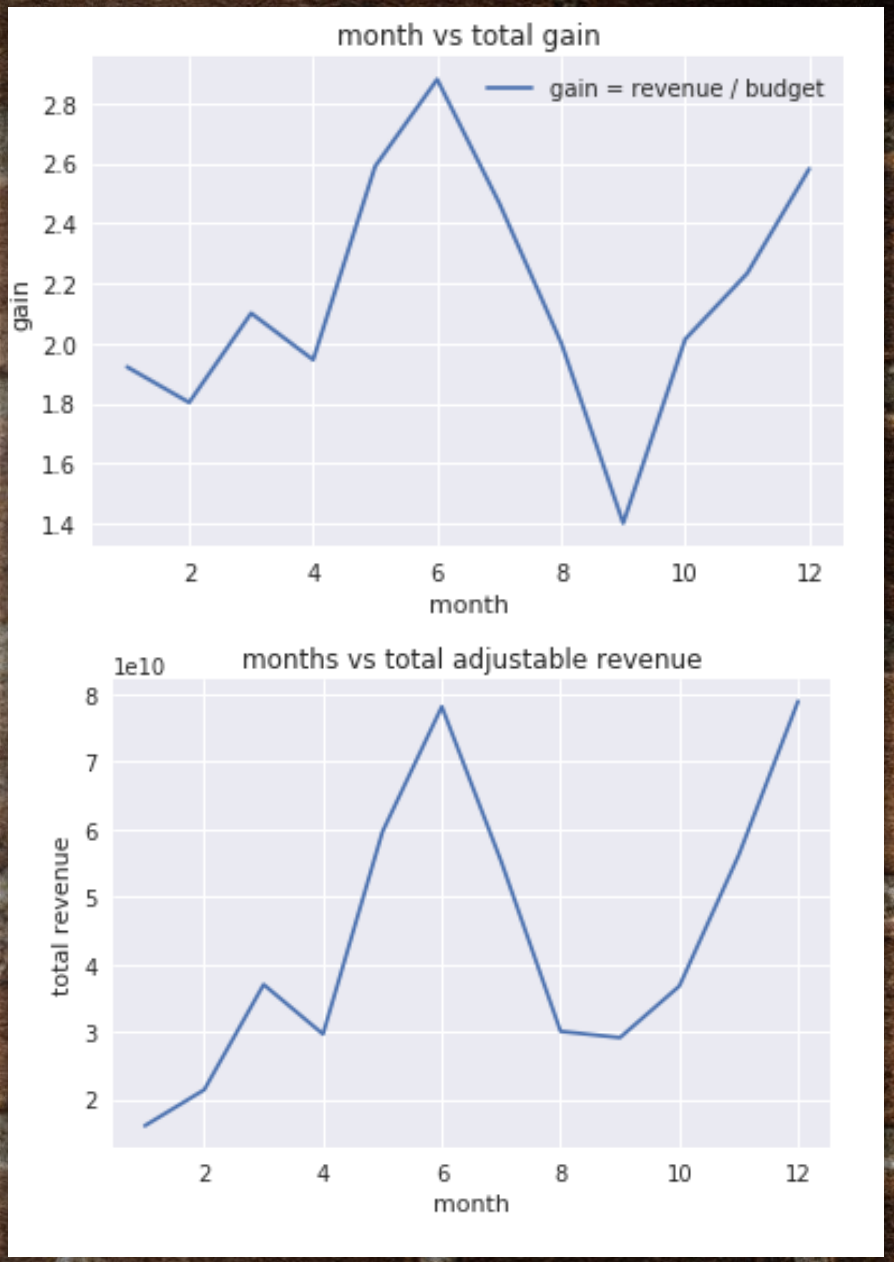
Don’t take the risk and publish your movie in month with low gain because you maybe will gain little revenue

It seems that the companies doesn’t know the best month because as you see

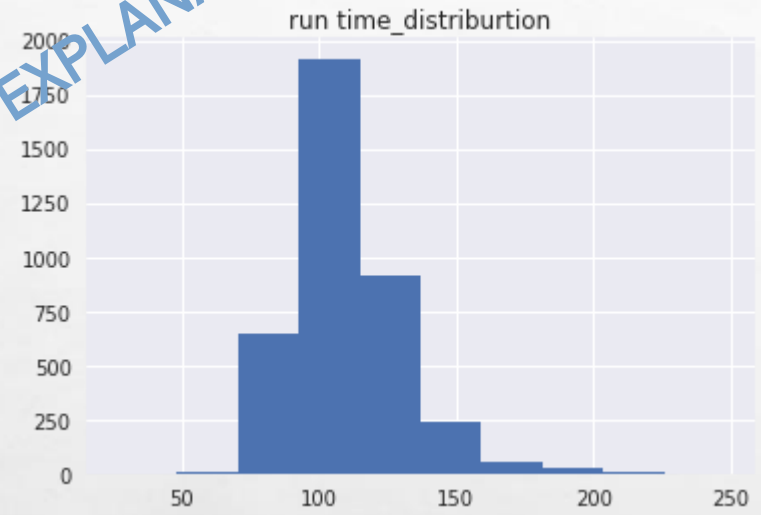
- the month with the maximum number of movies is 9
- the month with the minimum number of movies is 1



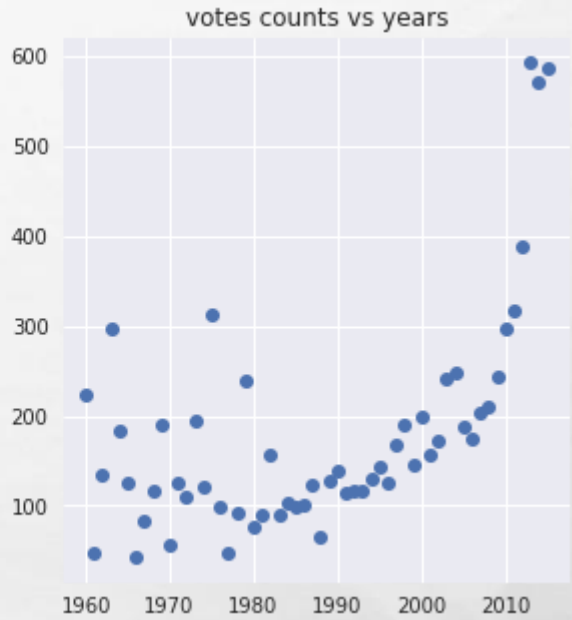
Gain mean the percentage of revenue to the total budget



RUN TIME EXPLANATION ?



VOTES EXPLANATION ?



Run time has no obvious relation with any variable .
Run time mean :- 109
Run time median :- 106

Its wired but over years number of votes increase but the average votes decrease.

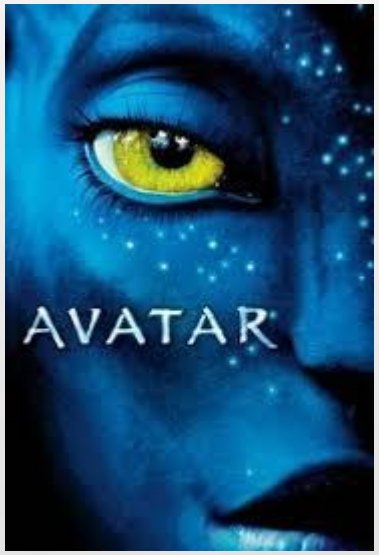
BEST MOVIES

HIGHER GAIN

THE ROCKY HORROR PICTURE SHOW

HIGHER REVENUES

AVATAR



HIGHER PROFIT

STAR WARS

HIGHER VOTE

THE SHAWSHANK REDEMPTION

Profit is → revenue – budget

WORST MOVIES

LOWER GAIN

DEATH DEFYING ACTS

LOWER REVENUES

BEST MAN DOWN

LOWER PROFIT

THE WARRIORS WAY

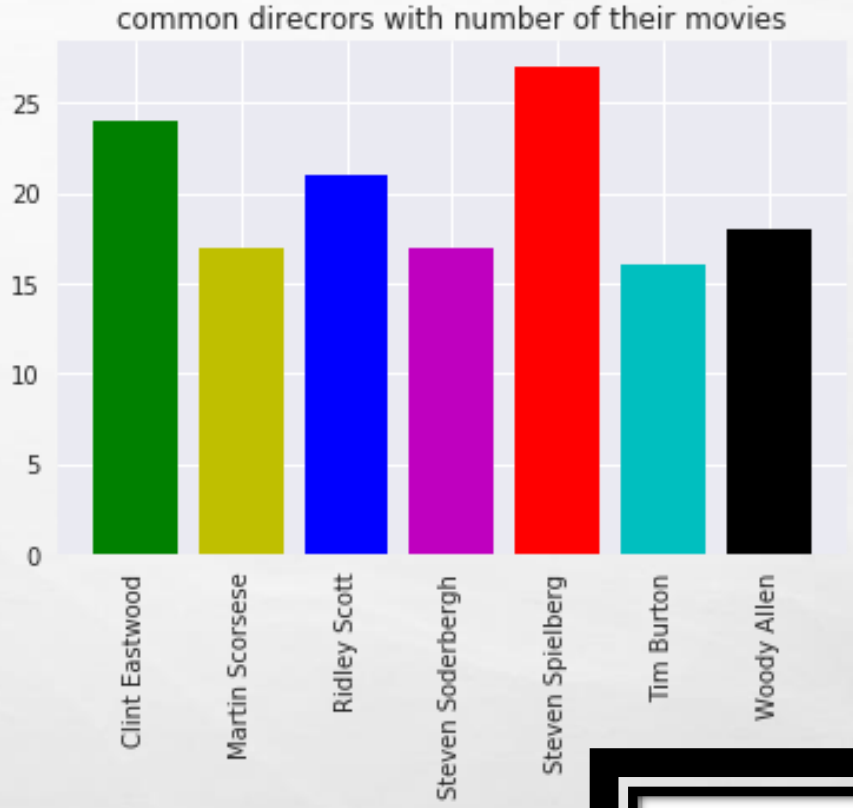
LOWER VOTE

FOODFIGHT!

the most popular director with more than15 movies .

The average results of there work

director	budget_adj	revenue_adj	gain	vote_average
Steven Spielberg	6.926672e+07	4.268546e+08	4.266667	6.9
Tim Burton	7.226958e+07	1.527780e+08	2.964176	6.6
Clint Eastwood	3.875860e+07	1.181567e+08	3.344524	6.7
Ridley Scott	7.089506e+07	1.147286e+08	1.628024	6.3
Martin Scorsese	5.132198e+07	1.082936e+08	2.232930	7.1
Steven Soderbergh	3.461341e+07	5.931901e+07	1.619699	6.3
Woody Allen	1.677039e+07	5.186416e+07	4.672915	6.8



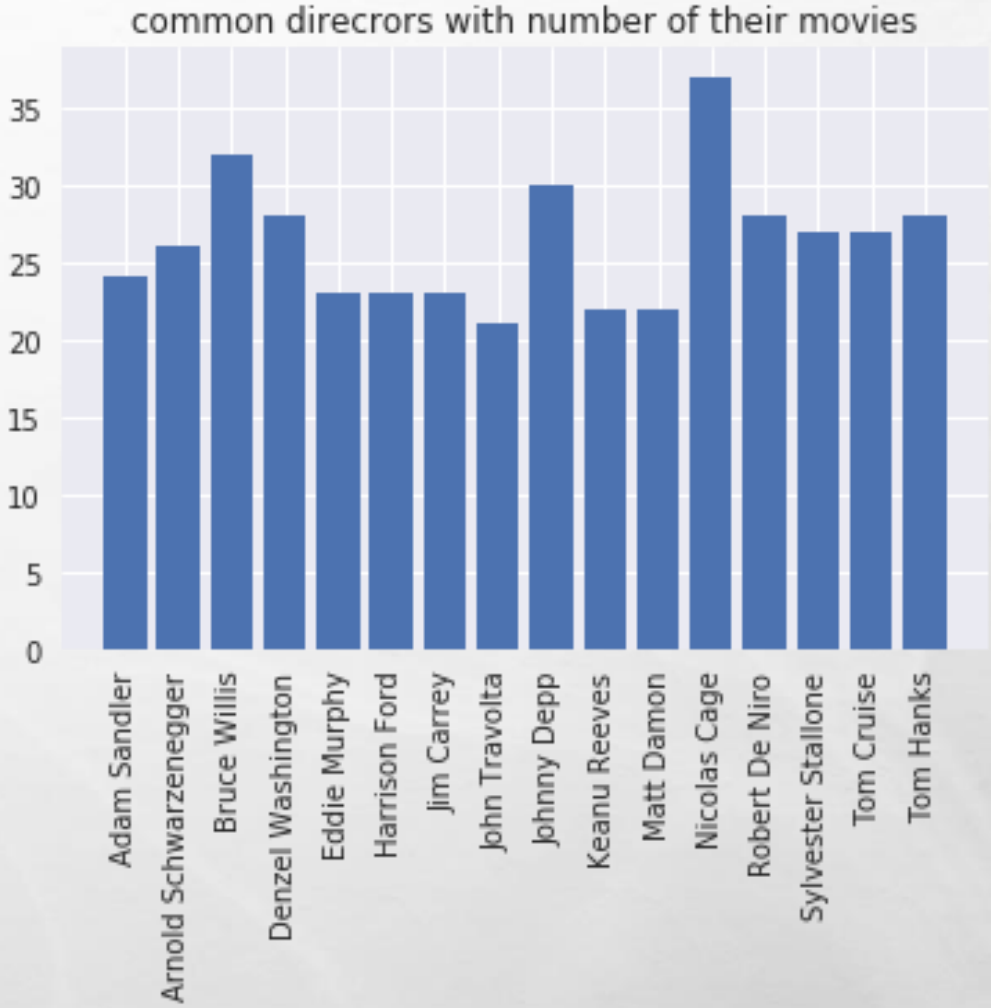
Steven Spielberg	27
Clint Eastwood	24
Ridley Scott	21
Woody Allen	18
Steven Soderbergh	17
Martin Scorsese	17
Tim Burton	16

The best director is Steven Spielberg

the most popular actors with more than 20 movies .

The average results of there work

main_hero	budget_adj	revenue_adj	gain	vote_average
John Travolta	6.125927e+07	2.173701e+08	9.352698	6.028571
Tom Hanks	7.686975e+07	3.456860e+08	5.042067	6.778571
Tom Cruise	9.380248e+07	3.656239e+08	4.949371	6.429630
Harrison Ford	7.978699e+07	3.530982e+08	4.369419	6.282609
Jim Carrey	6.361998e+07	2.151571e+08	4.195820	6.291304
Sylvester Stallone	6.719206e+07	2.119823e+08	4.103160	6.029630
Arnold Schwarzenegger	8.316739e+07	2.296522e+08	3.556022	5.853846
Eddie Murphy	7.131055e+07	1.966644e+08	3.515317	5.539130
Adam Sandler	6.509282e+07	1.847532e+08	3.089461	6.025000
Bruce Willis	7.306387e+07	1.967379e+08	2.972644	6.150000
Robert De Niro	5.513073e+07	1.111846e+08	2.788701	6.496429
Keanu Reeves	6.088136e+07	1.823470e+08	2.727395	6.204545
Matt Damon	6.461209e+07	1.477983e+08	2.291639	6.422727
Denzel Washington	6.265634e+07	1.307588e+08	2.158370	6.539286
Johnny Depp	9.534720e+07	2.300878e+08	2.008459	6.493333
Nicolas Cage	5.820798e+07	1.185813e+08	1.972249	5.856757



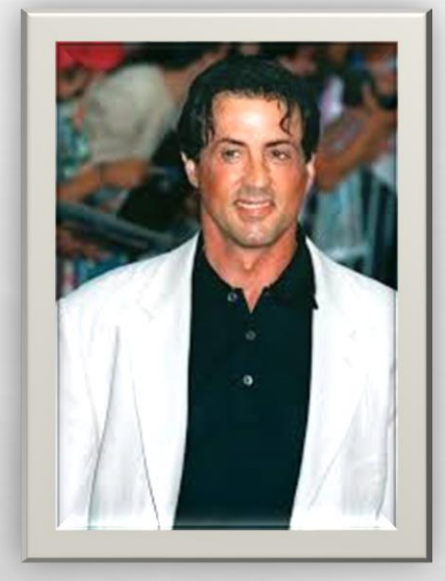
Its wired but actors that there name starts with tom are the best

director that love specific actor to work with

I studied the case that the actor work with the same director more than 5 times

main_hero	director	
Adam Sandler	Dennis Dugan	7
Clint Eastwood	Clint Eastwood	12
George Clooney	Steven Soderbergh	6
Johnny Depp	Tim Burton	6
Sylvester Stallone	Sylvester Stallone	6
Woody Allen	Woody Allen	10

It's wired but Sylvester Stallone is the director for the movies he act in



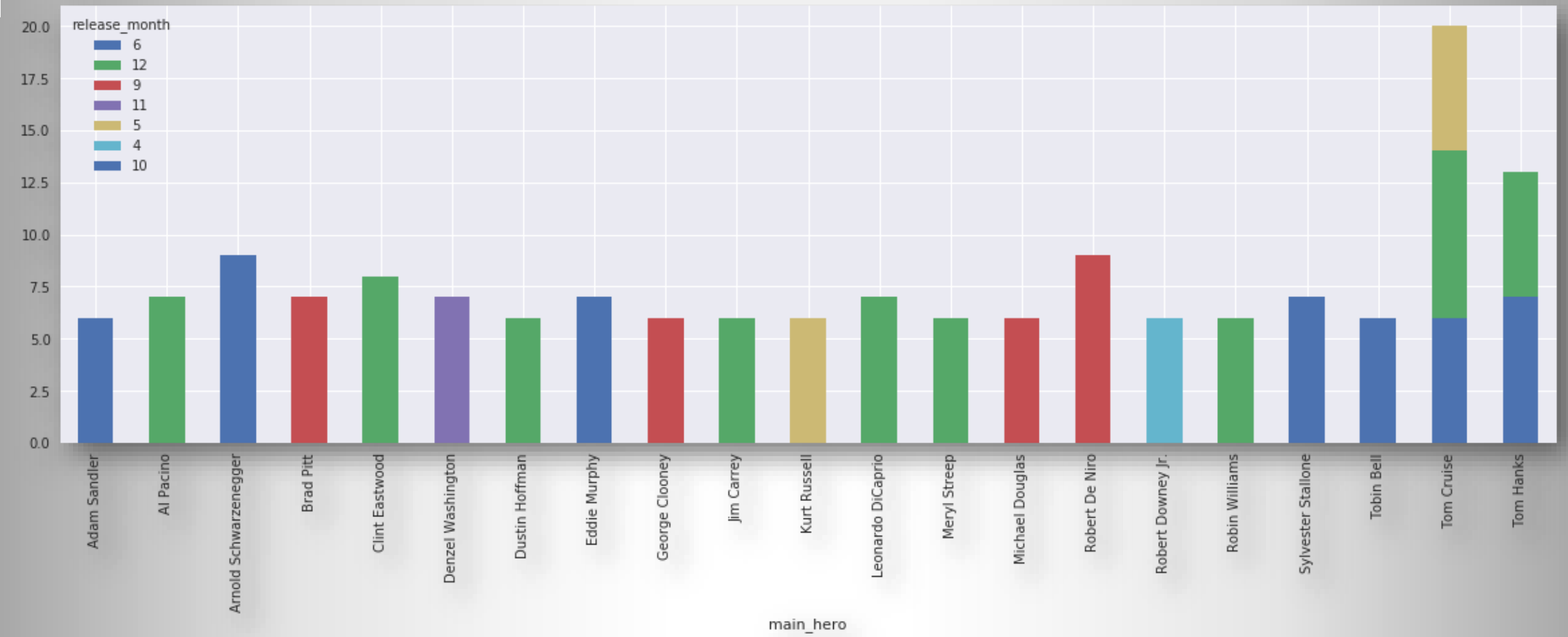
the most popular production company with more than 50 movies

Universal Pictures	282
Paramount Pictures	260
Columbia Pictures	177
Twentieth Century Fox Film Corporation	164
New Line Cinema	139
Walt Disney Pictures	116
Miramax Films	77
Warner Bros.	75
Columbia Pictures Corporation	66
Village Roadshow Pictures	66
DreamWorks SKG	59
TriStar Pictures	57

companies that work with the same actor and director more than 3 time

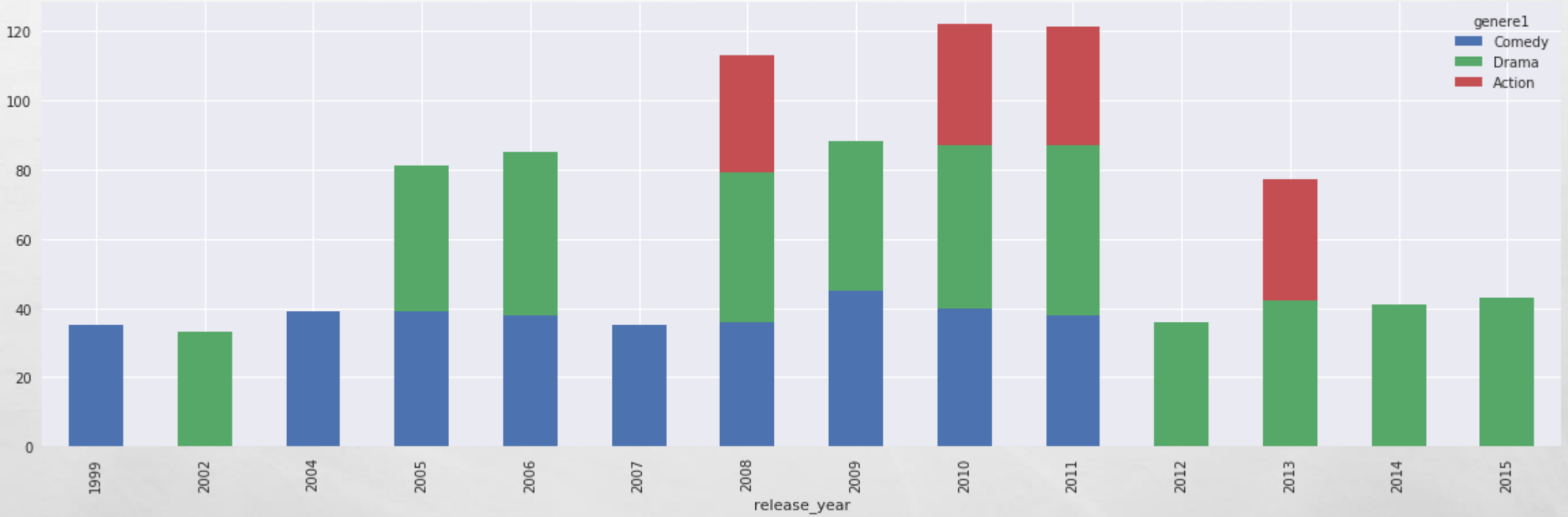
company1	main_hero	director	
Columbia Pictures	Adam Sandler	Dennis Dugan	5
Lions Gate Films	Tyler Perry	Tyler Perry	4
Lucasfilm	Harrison Ford	Steven Spielberg	4
Malpaso Productions	Clint Eastwood	Clint Eastwood	4
Silver Pictures	Mel Gibson	Richard Donner	5
Walt Disney Pictures	Johnny Depp	Gore Verbinski	4

That's because they gain more revenue with each other



There is a relation between the month of release and the actor there is actors that release their movie in the same month such as `Robert De Niro` love month 9

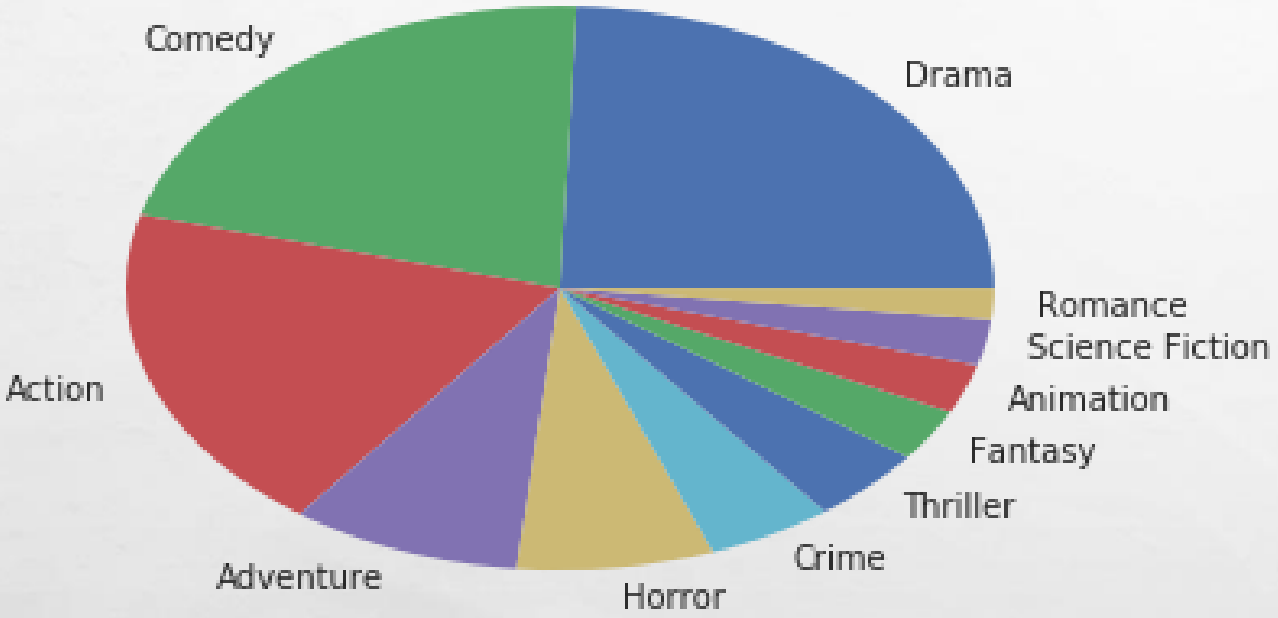
FIND THE POPULAR GENRE IN EACH YEAR



there is a relation as every year has a common movie category

I studied the genre which repeat more than 30 times at a year

common genre



Genres vs numbers

```
best_generes
the higher revenue is-->  Animation
the higher vote is -->   Documentary
the is higher profit -->  Animation
the higher gain is -->  TV Movie
```

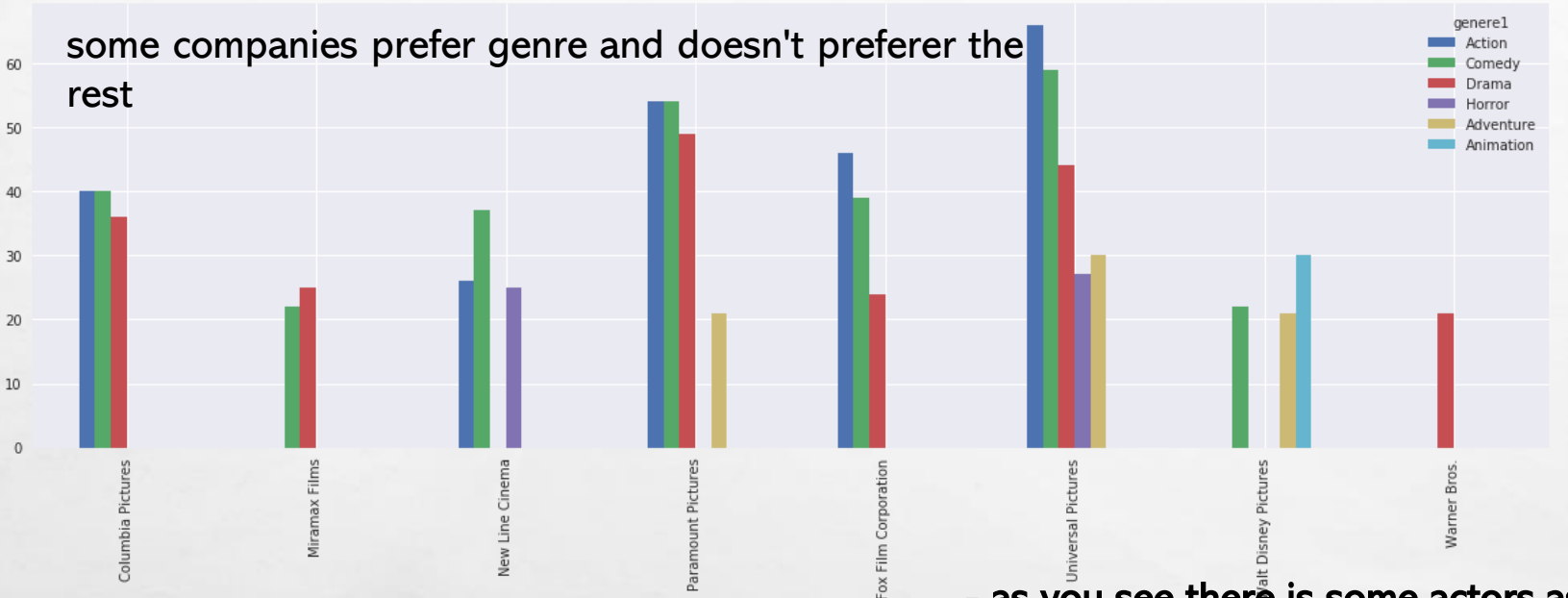
```
worest_movies
the lower revenue is-->  Documentary
the lower vote is -->   TV Movie
the is lower profit -->  Documentary
the lowe gain is -->  Mystery
```

the most common genre is drama

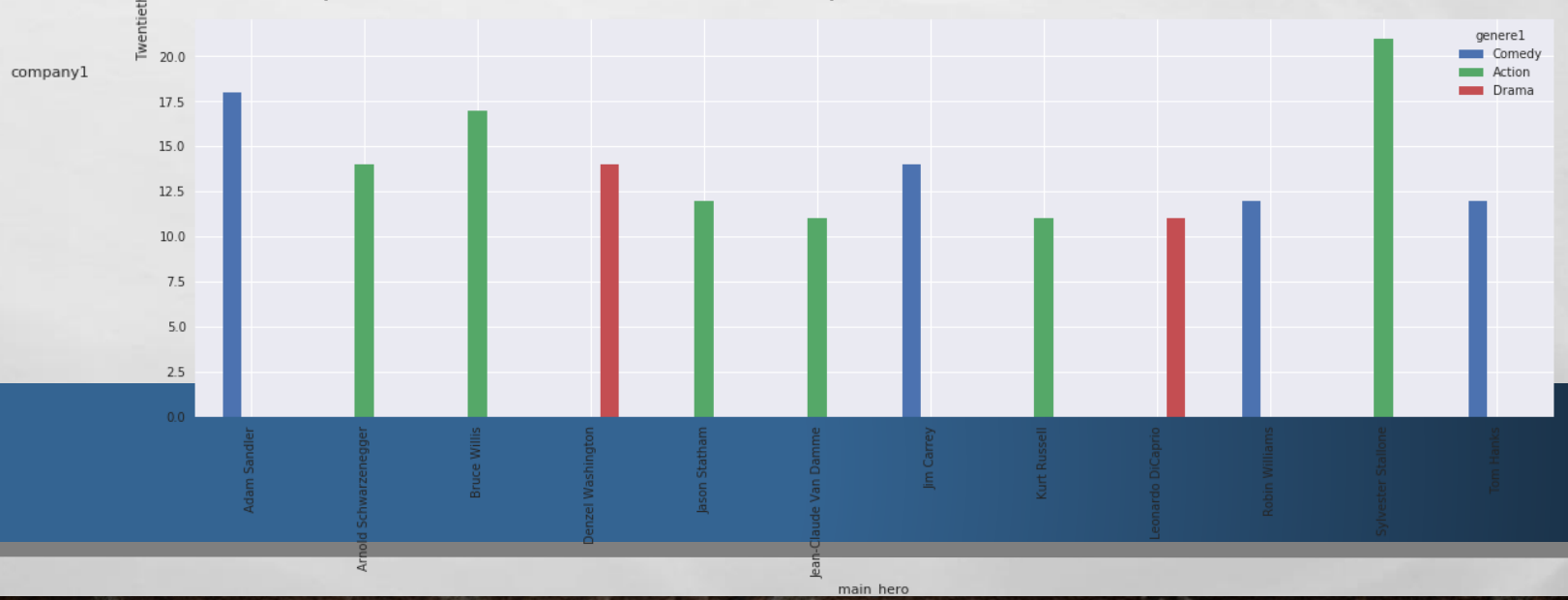
Note that genre that don’t exist in the pie chart has less than 50 movies

The relation between genre companies and directors

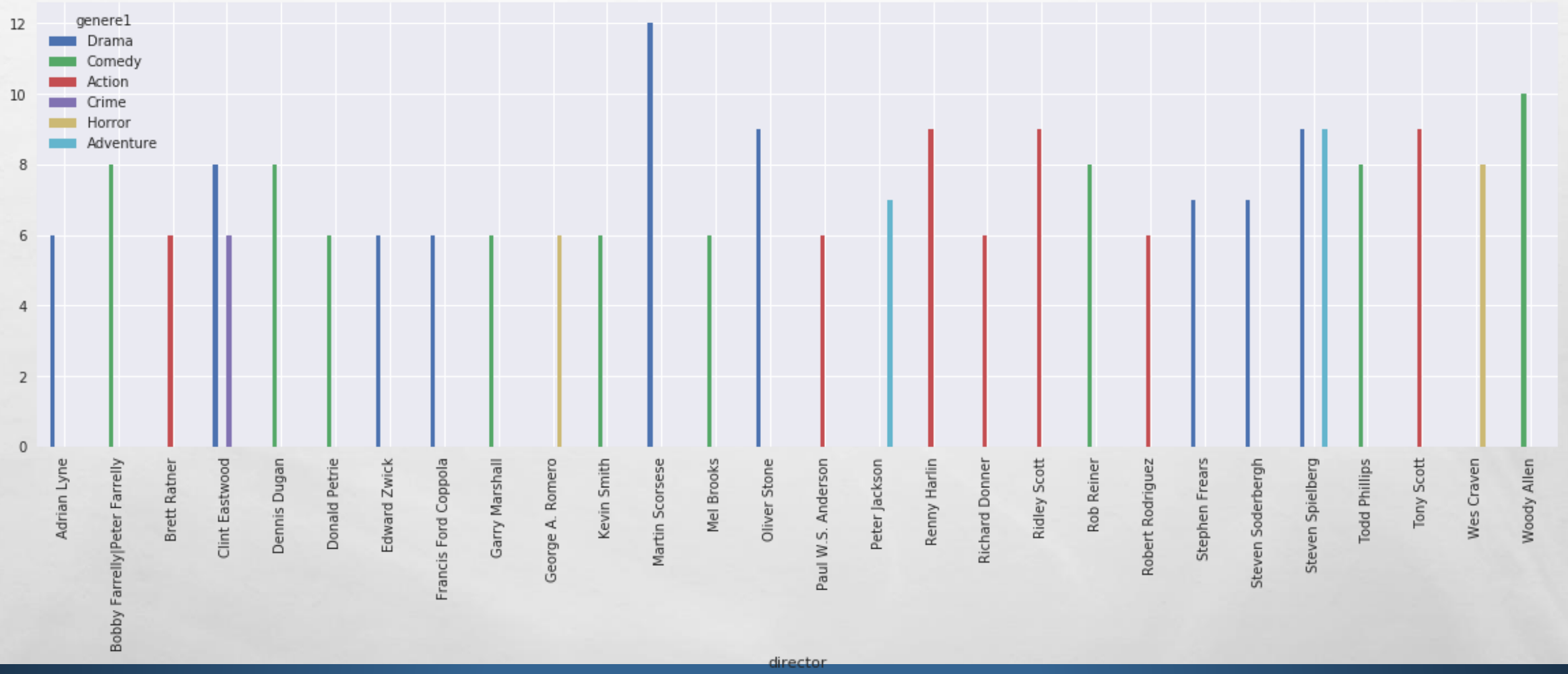
some companies prefer genre and doesn't preferer the rest



- as you see there is some actors as sylvester stallone prefer action and woody adam sandler prefer comedy



The relation between genre and directors



as you see there is some directors as martin Scorsese prefer drama and woody Allen prefer comedy

DATA CLEANING

WRONG DATA TYPE

I Converted release_date from string to time stamp

DUPLICATE

There is one duplicate raw so I delete it

THERE IS NOT IMPORTANT COLUMN

I will not use these column → {id ,

imdb_id , homepage , overview , tagline

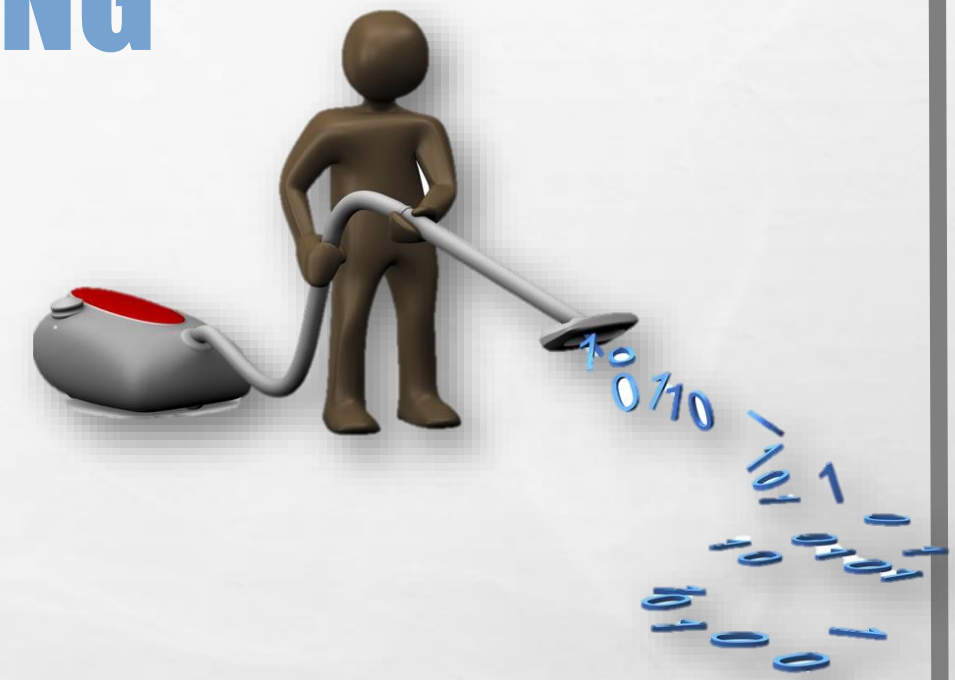
} so I droppes them

add extra column

release month , gain which equal to revenue / budjet / movie_hero

split columns

genere , production_company



INVALID DATA

There is some data that don't make sense such as minimum in revenue and budget is 0 , run_time maximum is to big


Thank you