TMDb movie Data Analysis project

Introduction

TMDb movie Data Analysis project about 10,000 movies collected from The Movie Database (TMDb). It includes some data such as releasing year, budget, revenue and ratings for those movies. In this analysis, I'll first clean the datasets to get the most accurate amount of data. Then after getting satisfied with the data, I'll visualize the analysis about some question and relationships between data columns. These questions:

- Is there a certain genre formula has highest rating?
- Is budget has a relation with rating?
- What is the relation between the movie revenue and movie rating?
- What is the movie genre has the highest revenue?

TMDB dataset started with shape of (10866, 21) with columns as shown below:

Data columns (total 21 columns):
id 10866 non-null int64
imdb_id 10856 non-null object
popularity 10866 non-null float64
budget 10866 non-null object
revenue 10866 non-null object
original_title 10866 non-null object
cast 10790 non-null object
homepage 2936 non-null object
director 10822 non-null object
tagline 8042 non-null object
keywords 9373 non-null object
overview 10862 non-null object
runtime 10866 non-null int64
genres 10843 non-null object
production_companies 9836 non-null object
release_date 10866 non-null object
vote_count 10866 non-null int64
vote_average 10866 non-null float64
release_year 10866 non-null int64
budget_adj 10866 non-null object
revenue_adj 10866 non-null object

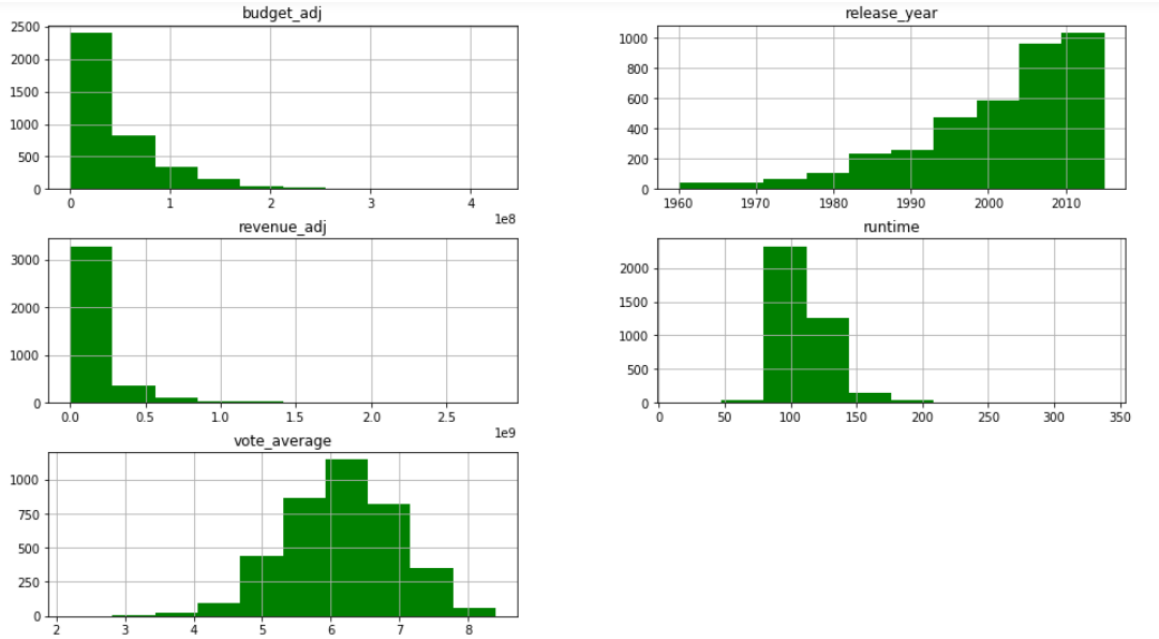
After cleaning TMDB dataset becomes with shape of (3807, 10) with columns as shown below:

Data columns (total 10 columns):
original_title 3807 non-null object
director 3807 non-null object
runtime 3807 non-null int64
genres 3807 non-null object
production_companies 3807 non-null object
release_date 3807 non-null object
vote_average 3807 non-null float64
release_year 3807 non-null int64
budget_adj 3807 non-null int64
revenue_adj 3807 non-null int64

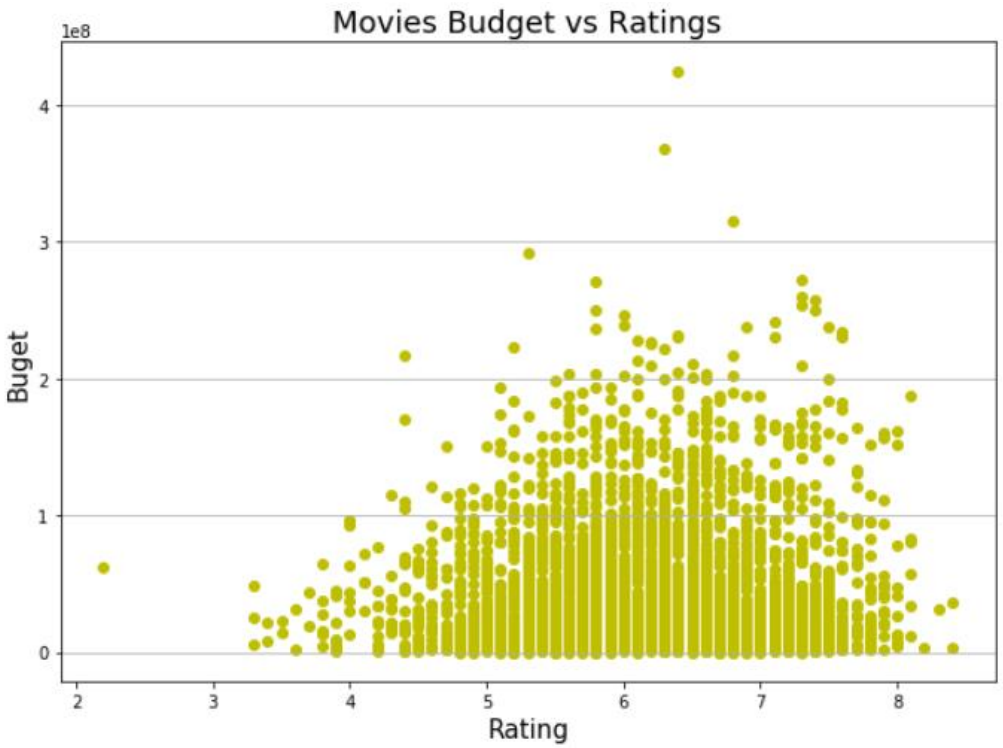
	original_title	director	runtime	genres	production_companies	release_date	vote_average	release_year	budget_adj	revenue_adj
0	Jurassic World	Colin Trevorrow	124	Action|Adventure|Science Fiction|Thriller	Universal Studios|Amblin Entertainment|Legenda...	6/9/2015	6.5	2015	$137,999,939.28	$1,392,445,892.52
1	Mad Max: Fury Road	George Miller	120	Action|Adventure|Science Fiction|Thriller	Village Roadshow Pictures|Kennedy Miller Produ...	5/13/2015	7.1	2015	$137,999,939.28	$348,161,292.49
2	Insurgent	Robert Schwentke	119	Adventure|Science Fiction|Thriller	Summit Entertainment|Mandeville Films|Red Wago...	3/18/2015	6.3	2015	$101,199,955.47	$271,619,025.41
3	Star Wars: The Force Awakens	J.J. Abrams	136	Action|Adventure|Science Fiction|Fantasy	Lucasfilm|Truenorth Productions|Bad Robot	12/15/2015	7.5	2015	$183,999,919.04	$1,902,723,129.80
4	Furious 7	James Wan	137	Action|Crime|Thriller	Universal Pictures|Original Film|Media Rights ...	4/1/2015	7.3	2015	$174,799,923.09	$1,385,748,801.47

Data Analysis with visualization

Histograms below show that number of movies increases with time "per year". Cinema industry is flourishing over the years, which makes us wonder why this flourishing is fast?

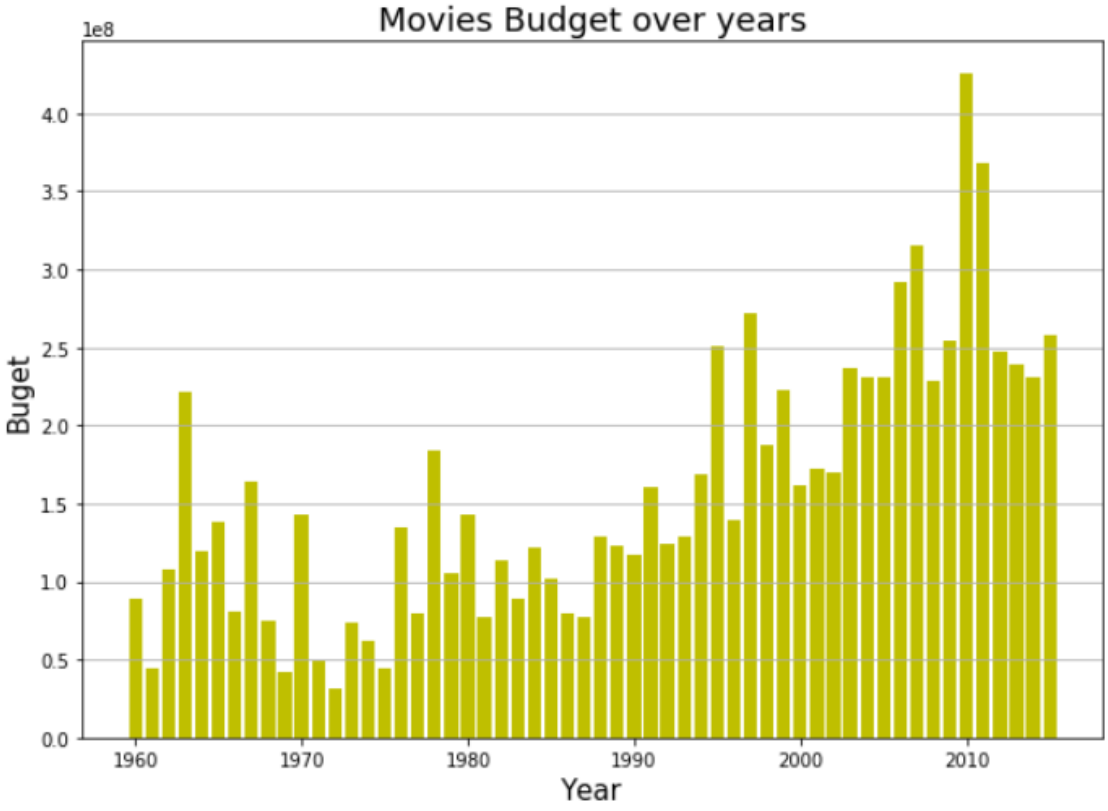


Research Question 1 (Is there a certain genre formula has highest rating?)



The initial impression that the ratings have no direct relation with movie budget. So let's be more specified and measure over some fixed values.

#let's see if the movie budget increases over years?



As shown, there's a budget increasing over the years especially since late 90's and overall budget increases since 2000. Not necessary it's a continuous increment but overall there's a budget jumping.

>> if the budget has a relation with the rate over cinema flourishing years?

	vote_average	budget_adj	release_year
0	6.5	137999939	2015
89	5.9	45999979	2015
118	6.3	11039995	2015
120	5.3	22999989	2015
121	6.4	50599977	2015

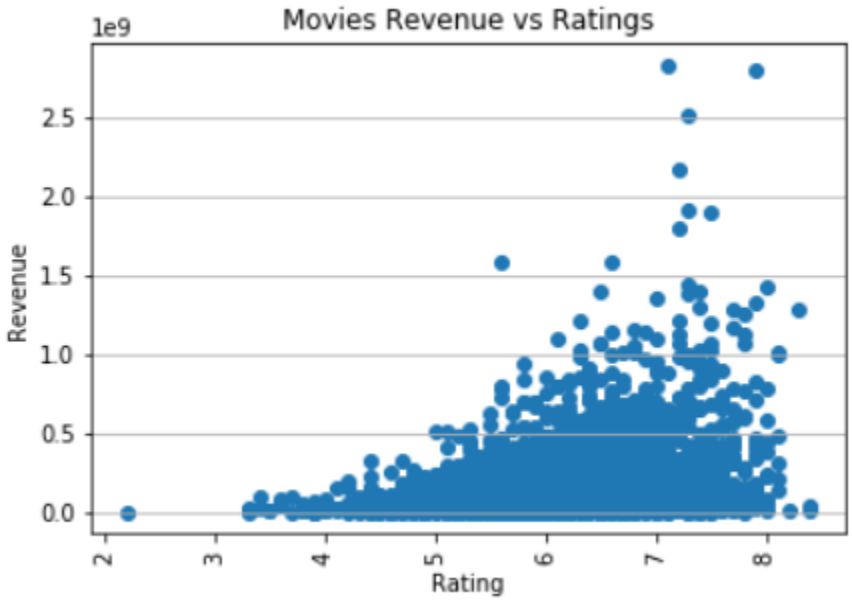
>> what if over a movie series, budget and rating? "if there as some fixed variables such as director, etc.."

	original_title	director	runtime	genres	production_companies	release_date	vote_average	release_year	budget_adj	revenue_adj
634	The Hobbit: The Battle of the Five Armies	Peter Jackson	144	Adventure|Fantasy	WingNut Films|New Line Cinema|3Foot7|Metro-Gol...	12/10/2014	7.1	2014	230272762	879752289
5431	The Hobbit: The Desolation of Smaug	Peter Jackson	161	Adventure|Fantasy	WingNut Films|New Line Cinema|Metro-Goldwyn-Ma...	12/11/2013	7.6	2013	234008338	897094365
4367	The Hobbit: An Unexpected Journey	Peter Jackson	169	Adventure|Fantasy|Action	WingNut Films|New Line Cinema|Metro-Goldwyn-Ma...	11/26/2012	6.9	2012	237436070	965893322

Conclusion 1: No proportional relation between movies budget & rating over a certain time of years also over a certain movie series "although production_company & director are the same" which indicates that content is more crucial.

In 90's Titanic budget was huge compared to cinema industry at this time. While there was a budget flourish started in 2003 investing in Terminator movie and this flourish continued investing in good movies like King Kong and a peak of The Warrior's Way although it has an average rating.

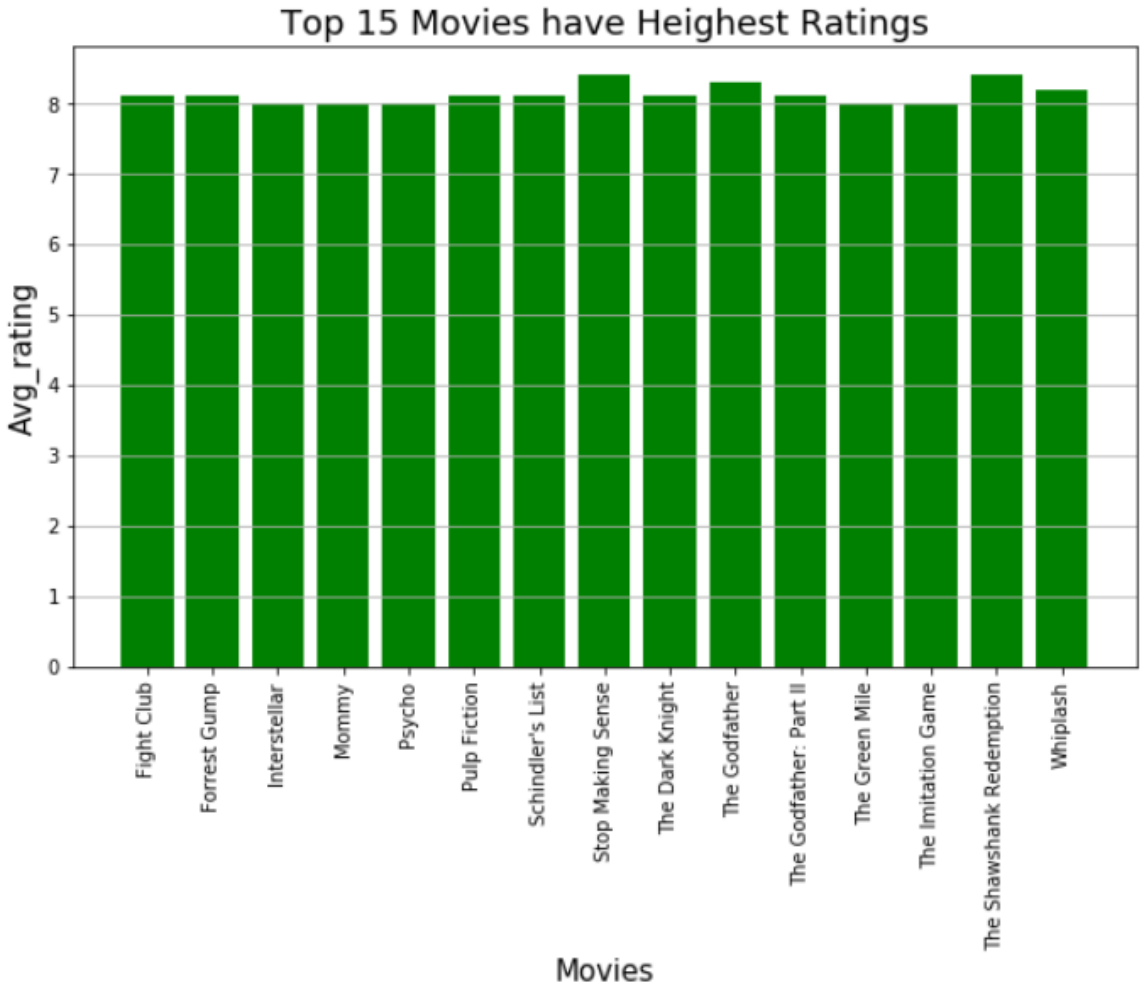
Research Question 2 (What is the relation between the movie revenue and movie rating?)



Conclusion 2: here's a weak linearity between movies rating and revenue_adj due to outliers. ¶

Research Question 3 (Is there a certain genre formula has highest rating?)

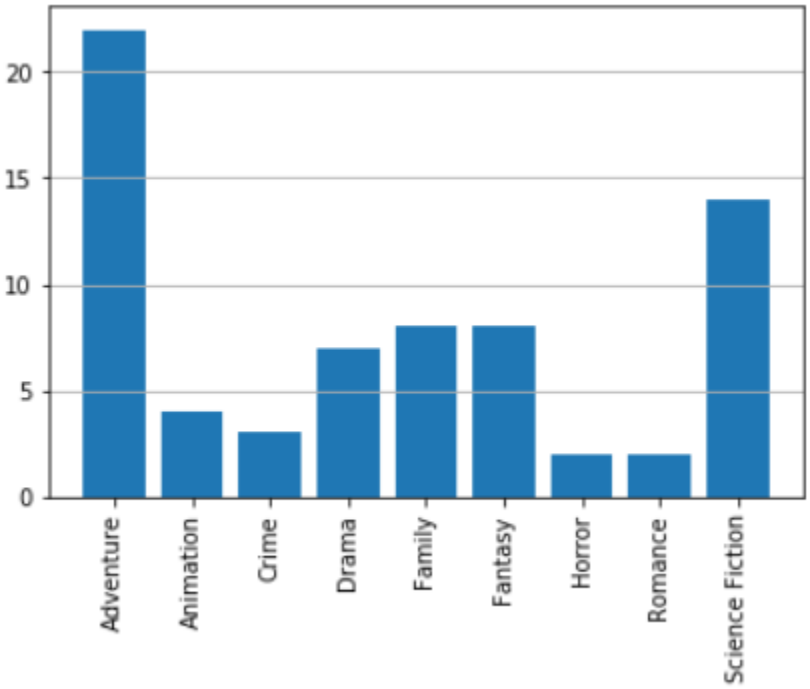
rating more than 7 :



"Stop Making Sense" got the highest rate in Documentary movies and "The Shawshank Redemption" as Drama & Crime movie

Conclusion 3: Drama genre of movies has the highest ratings through the years.

Research Question 4 (What is the movie genre has the highest revenue?)



Conclusion 4: Adventure movie genre has the highest revenue over the time, then comes science fiction at the second place.

Conclusions

About budget: In 90's Titanic budget was huge compared to cinema industry at this time. While there was a budget flourishing started in 2003 investing in Terminator movie and this flourish continued investing in good movies such as King Kong and a peak of The Warrior's Way although it has an average rating. Also, there is no proportional relation between movies budget & rating over a certain time of years also over a certain movie series "although production_company & director are the same" which indicates that content is more crucial.

About ratings: Good movie rating not only has no relation with buget but There's a weak linearity between movies rating and revenue_adj due to outliers. Drama genre of movies has the highest ratings through the years.

About movies genres: Adventure movie genre has the highest revenue over the time, then comes science fiction at the second place.