IMDB Rating Prediction

I. Overview data

First, we will look at the structure of the data and interpret the meaning of feature columns, and then evaluate some problems in the data. Data includes 5043 rows of different movies and 28 columns that are the information related to the movies.

color	color movies or black and white movies
director_name	
num_critic_for_reviews	number of critics. The popularity of a movie in social network can be largely affected by the critics
duration	the time length of the movies in minutes
director_facebook_likes	number of likes from the facebook of the director
actor_1_facebook_likes	number of like from the facebook of actor/actress that has the highest number of like for all cast members
actor_2_facebook_likes	number of like from the facebook of actor/actress that has the second lagest number of like for all cast members
actor_3_facebook_likes	number of like from the facebook of actor/actress that has the third lagest number of like for all cast members
cast_total_facebook_likes	number of likes from the facebook of all the cast members.
movie_facebook_likes	number of like from facebook of the movie
actor_1_name	name of actor/actress that has the highest number of like from facebook for all cast members.
actor_2_name	name of actor/actress that has the second largest number of like from facebook for all cast members.
actor_3_name	name of actor/actress that has the third largest number of like from facebook for all cast members.
gross	gross earning of the movie
genres	movies genres
movie_title	
num_voted_users	number of user that voted for the movie
facenumber_in_poster	number of human faces in movie posters
plot_keywords	keywords from the movie content
movie_imdb_link	the link of the movie on imdb website
num_user_for_reviews	number of users that write review for the movie
language	language in the movie
country	country which the movie was produced
content_rating	which age group is suitable to watch, such as R, PG, PG-13, etc.
budget	budget to produce the movie
title_year	the year that movie was produced
imdb_score	imdb rating (targer variable)
aspect_ratio	the widescreen of the movie (16: 9 for example)

There are some data problems that we need to handle:

1. Columns with object types: these columns are text or category type, not numerical, so we need to transform it into dummy variables.
2. Object columns with many different values, so we need to group these values to avoid creating to many dummy variables.
3. Missing data (21 columns).
4. Remove duplicates.

Data has 5043 rows and 28 columns.

Number of column type:

float64:	13
object:	12
int64:	3

Remove 45 duplicates.

There are 21 columns that have missing values:

Column	Number of missing values	%
gross	884	17.5
budget	492	9.8
aspect_ratio	329	6.5
content_rating	303	6
plot_keywords	153	3
title_year	108	2.1
director_name	104	2.1
director_facebook_likes	104	2.1
num_critic_for_reviews	50	1
country	5	0.1
actor_3_name	23	0.5
actor_3_facebook_likes	23	0.5
num_user_for_reviews	21	0.4
color	19	0.4
duration	15	0.3
facenumber_in_poster	13	0.3
actor_2_name	13	0.3
actor_2_facebook_likes	13	0.3
language	12	0.2
actor_1_name	7	0.1
actor_1_facebook_likes	7	0.1

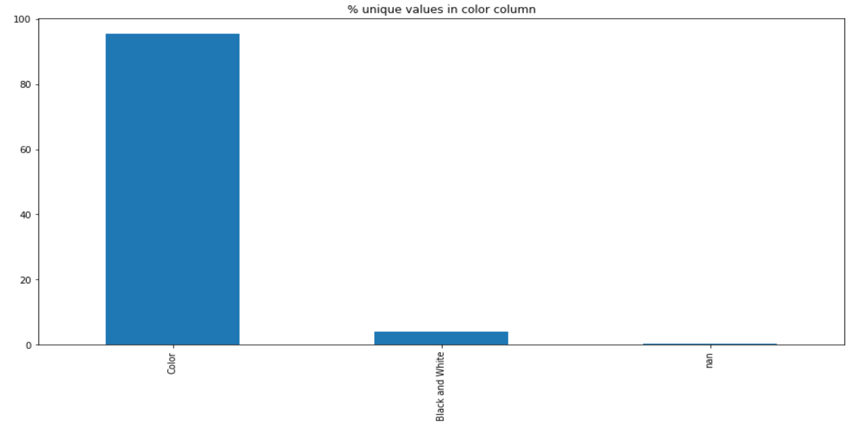
II. Data Visualization

Number of unique values:

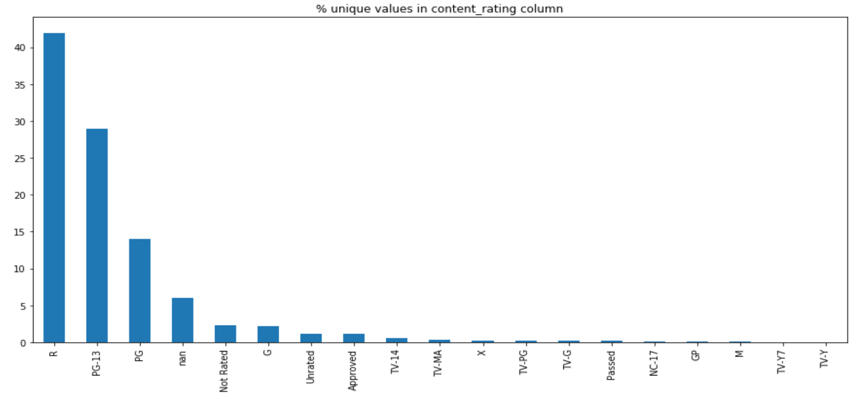
Column	Unique value
actor_3_facebook_likes	906
actor_1_facebook_likes	878
actor_2_facebook_likes	917
num_user_for_reviews	954
actor_1_name	2097
director_name	2398
actor_2_name	3032
actor_3_name	3521
cast_total_facebook_likes	3978
gross	4035
plot_keywords	4760
num_voted_users	4826
movie_title	4917
movie_imdb_link	4919

Column	Unique value
color	2
content_rating	18
facenumber_in_poster	19
aspect_ratio	22
language	47
country	65
imdb_score	78
title_year	91
duration	191
director_facebook_likes	435
budget	439
num_critic_for_reviews	528
movie_facebook_likes	876
genres	914

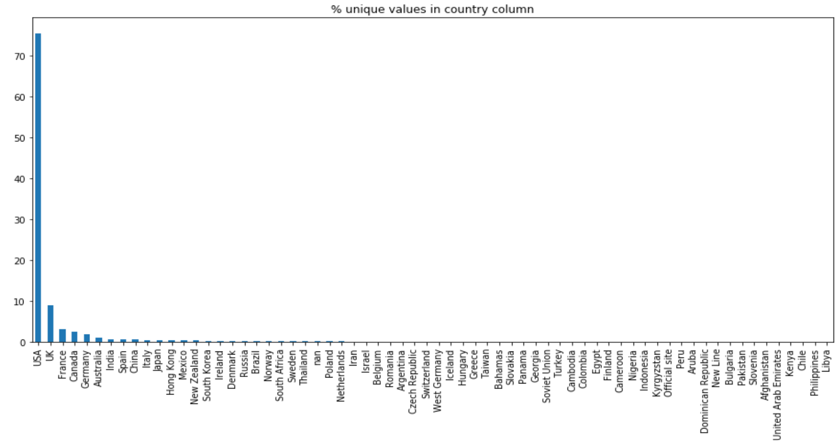
- Check columns with unique values < 100.



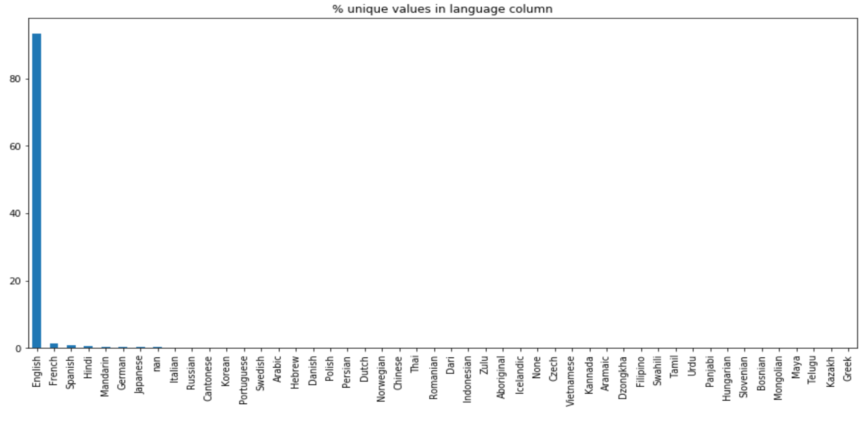
% unique values in color column



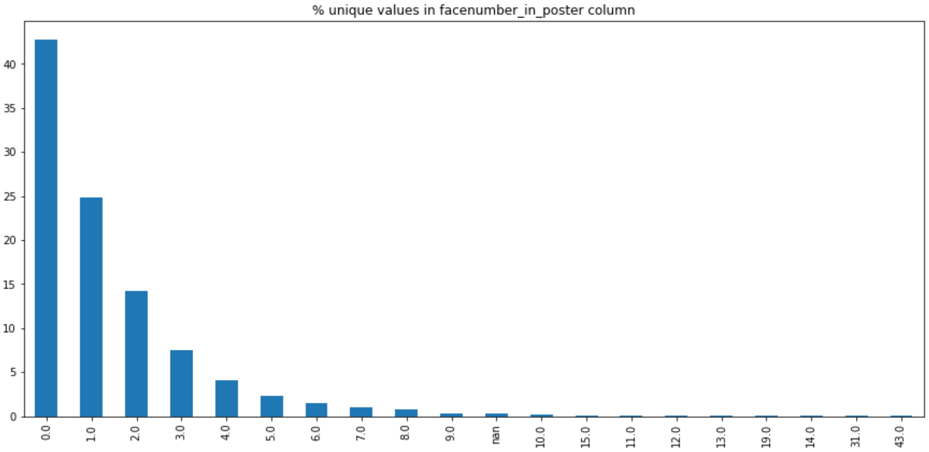
% unique values in content_rating column

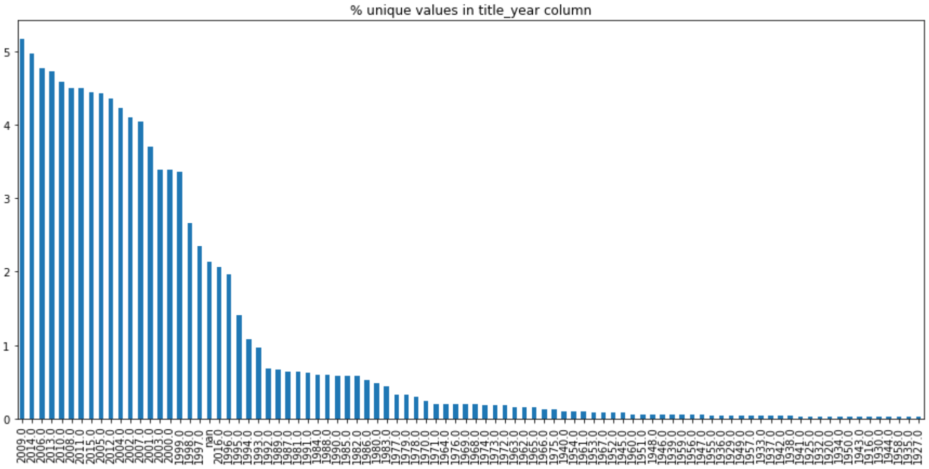


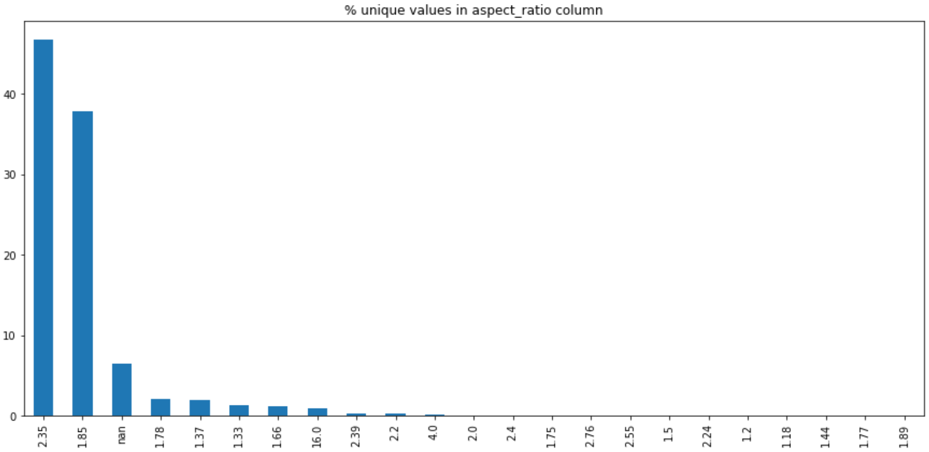
% unique values in country column

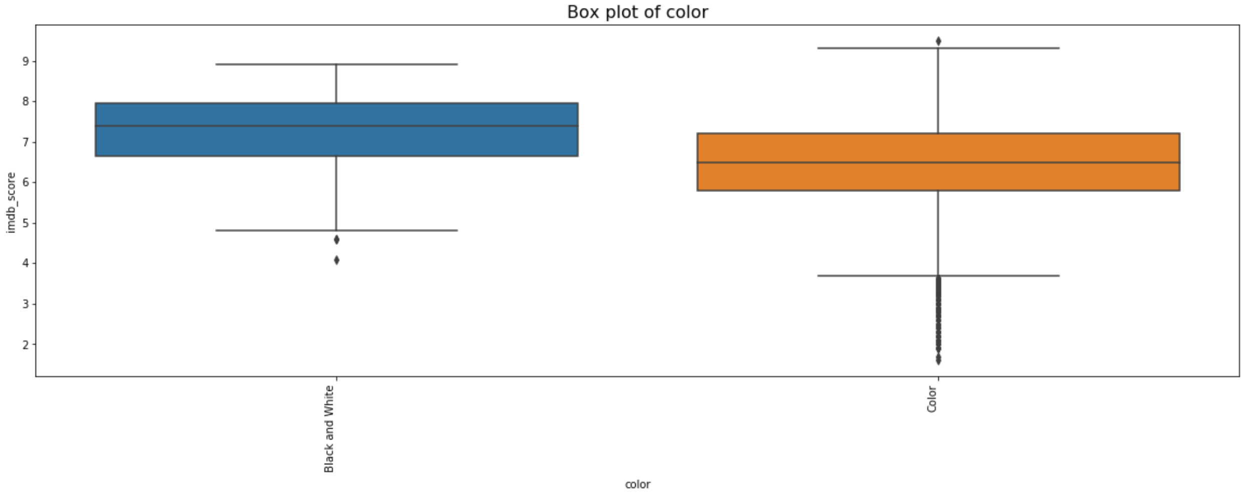


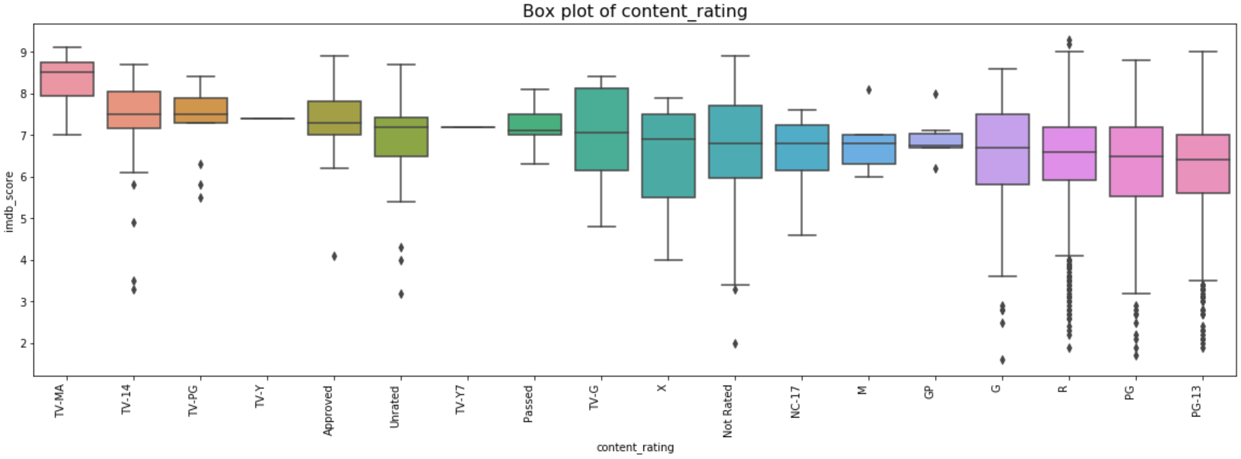
% unique values in language column

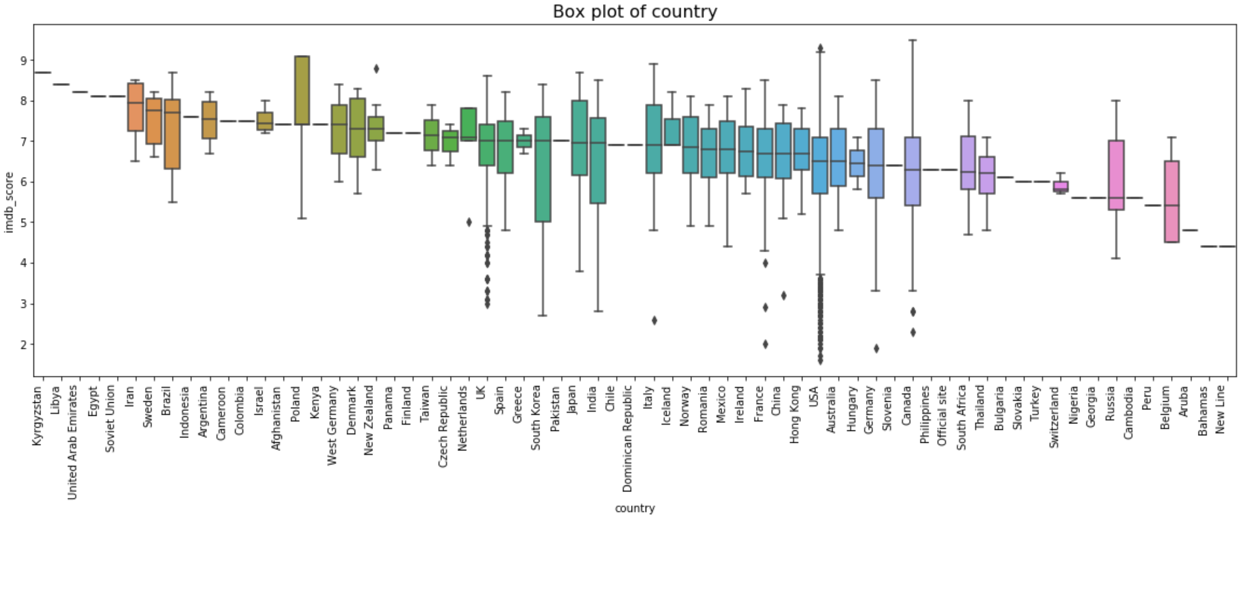

% unique values in facenumber_in_poster column
0
5
10
15
20
25
30
35
40
0.0
1.0
2.0
3.0
4.0
5.0
6.0
7.0
8.0
9.0
nan
10.0
15.0
11.0
12.0
13.0
19.0
14.0
31.0
43.0

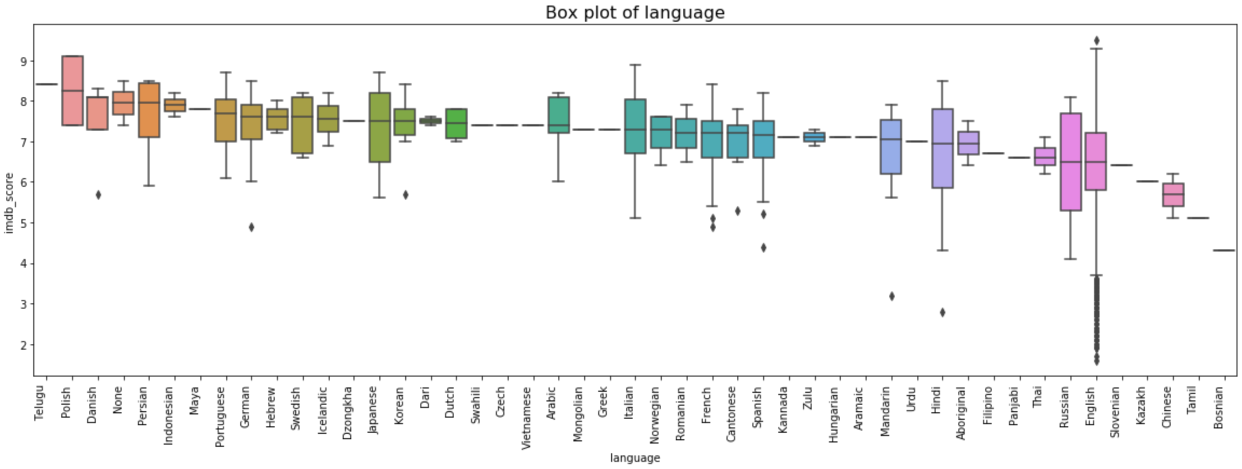

% unique values in title_year column
0
1
2
3
4
5
2009.0
2014.0
2006.0
2013.0
2010.0
2008.0
2011.0
2015.0
2005.0
2012.0
2004.0
2002.0
2007.0
2001.0
2003.0
2000.0
1999.0
1998.0
1997.0
nan
2016.0
1996.0
1995.0
1994.0
1993.0
1992.0
1989.0
1987.0
1981.0
1991.0
1984.0
1988.0
1990.0
1985.0
1982.0
1986.0
1980.0
1983.0
1977.0
1979.0
1978.0
1970.0
1971.0
1964.0
1976.0
1969.0
1968.0
1974.0
1973.0
1972.0
1963.0
1962.0
1965.0
1966.0
1975.0
1940.0
1954.0
1961.0
1953.0
1967.0
1952.0
1945.0
1960.0
1951.0
1948.0
1946.0
1939.0
1959.0
1956.0
1947.0
1955.0
1936.0
1929.0
1949.0
1957.0
1933.0
1937.0
1942.0
1938.0
1941.0
1925.0
1932.0
1920.0
1934.0
1950.0
1943.0
1916.0
1930.0
1944.0
1958.0
1935.0
1927.0

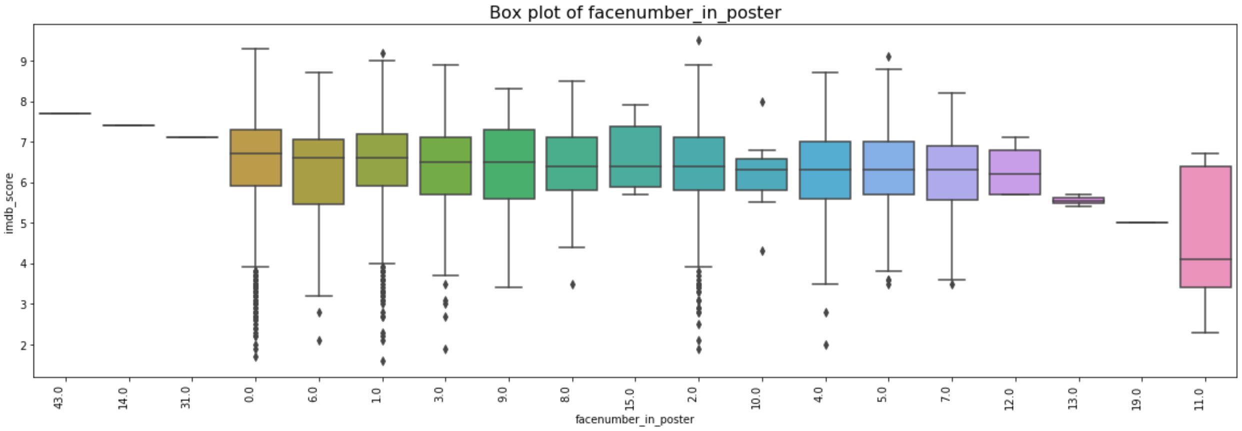

% unique values in aspect_ratio column
0
10
20
30
40
2.35
1.85
nan
1.78
1.37
1.33
1.66
16.0
2.39
2.2
4.0
2.0
2.4
1.75
2.76
2.55
1.5
2.24
1.2
1.18
1.44
1.77
1.89

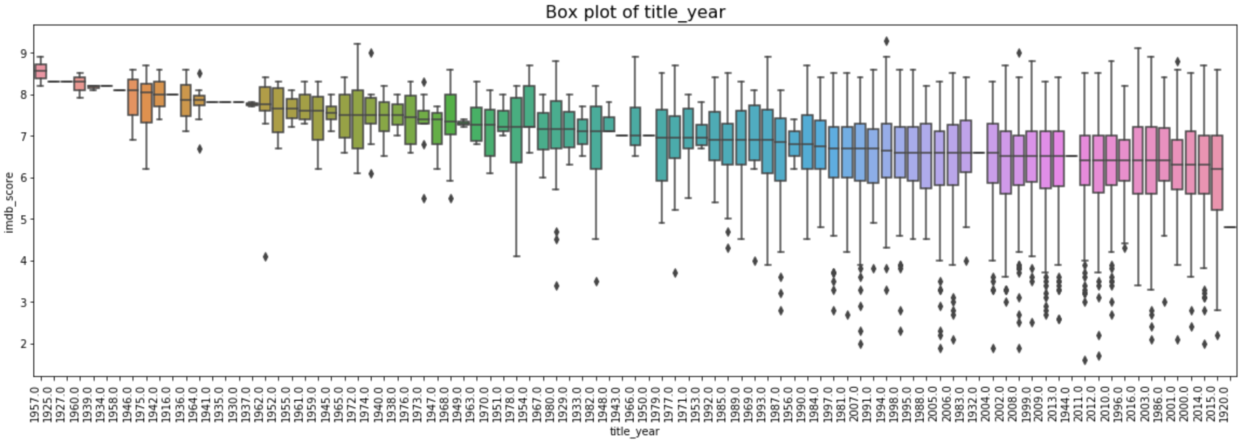

Box plot of color
imdb_score
9
8
7
6
5
4
3
2
Black and White
Color
color

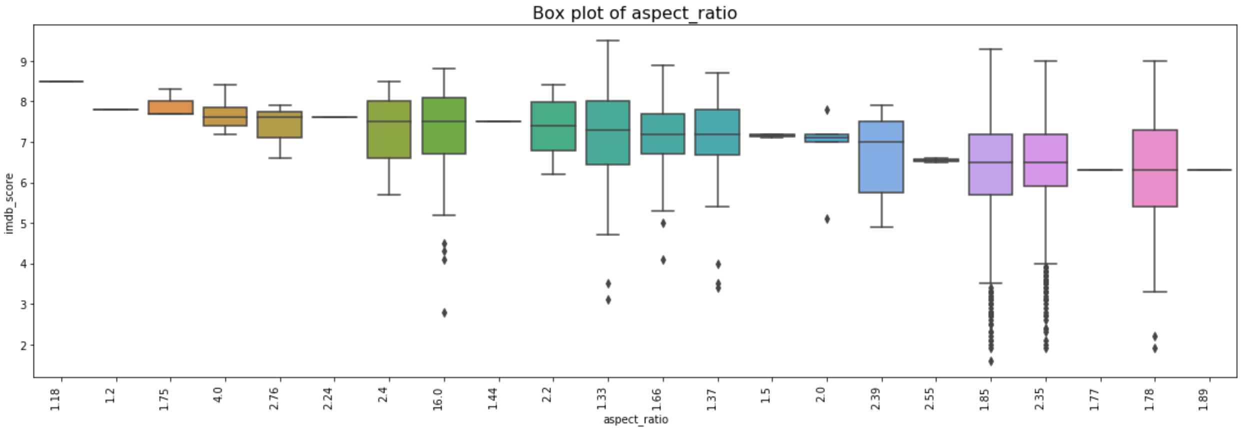

Box plot of content_rating
imdb_score
9
8
7
6
5
4
3
2
TV-MA
TV-14
TV-PG
TV-Y
Approved
Unrated
TV-Y7
Passed
TV-G
X
Not Rated
NC-17
M
GP
G
R
PG
PG-13
content_rating


Box plot of country
imdb_score
9
8
7
6
5
4
3
2
Kyrgyzstan
Libya
United Arab Emirates
Egypt
Soviet Union
Iran
Sweden
Brazil
Indonesia
Argentina
Cameroon
Colombia
Israel
Afghanistan
Poland
Kenya
West Germany
Denmark
New Zealand
Panama
Finland
Taiwan
Czech Republic
Netherlands
UK
Spain
Greece
South Korea
Pakistan
Japan
India
Chile
Dominican Republic
Italy
Iceland
Norway
Romania
Mexico
Ireland
France
China
Hong Kong
USA
Australia
Hungary
Germany
Slovenia
Canada
Philippines
Official site
South Africa
Thailand
Bulgaria
Slovakia
Turkey
Switzerland
Nigeria
Georgia
Russia
Cambodia
Peru
Belgium
Aruba
Bahamas
New Line
country


Box plot of language
imdb_score
9
8
7
6
5
4
3
2
Telugu
Polish
Danish
None
Persian
Indonesian
Maya
Portuguese
German
Hebrew
Swedish
Icelandic
Dzongkha
Japanese
Korean
Dari
Dutch
Swahili
Czech
Vietnamese
Arabic
Mongolian
Greek
Italian
Norwegian
Romanian
French
Cantonese
Spanish
Kannada
Zulu
Hungarian
Aramaic
Mandarin
Urdu
Hindi
Aboriginal
Filipino
Panjabi
Thai
Russian
English
Slovenian
Kazakh
Chinese
Tamil
Bosnian
language


Box plot of facenumber_in_poster
imdb_score
9
8
7
6
5
4
3
2
43.0
14.0
31.0
0.0
6.0
1.0
3.0
9.0
8.0
15.0
2.0
10.0
4.0
5.0
7.0
12.0
13.0
19.0
11.0
facenumber_in_poster


Box plot of title_year
imdb_score
9
8
7
6
5
4
3
2
1957.0
1925.0
1927.0
1960.0
1939.0
1934.0
1958.0
1946.0
1975.0
1942.0
1916.0
1936.0
1964.0
1941.0
1935.0
1930.0
1937.0
1962.0
1952.0
1955.0
1961.0
1959.0
1945.0
1965.0
1972.0
1974.0
1940.0
1938.0
1976.0
1973.0
1947.0
1968.0
1949.0
1963.0
1970.0
1951.0
1978.0
1954.0
1967.0
1980.0
1929.0
1933.0
1982.0
1948.0
1943.0
1966.0
1950.0
1979.0
1977.0
1971.0
1953.0
1992.0
1985.0
1989.0
1969.0
1993.0
1987.0
1956.0
1990.0
1984.0
1997.0
1981.0
2007.0
1991.0
1994.0
1998.0
1995.0
1988.0
2005.0
2006.0
1983.0
1932.0
2004.0
2002.0
2008.0
1999.0
2009.0
2013.0
1944.0
2011.0
2012.0
2010.0
1996.0
2016.0
2003.0
1986.0
2001.0
2000.0
2014.0
2015.0
1920.0
title_year

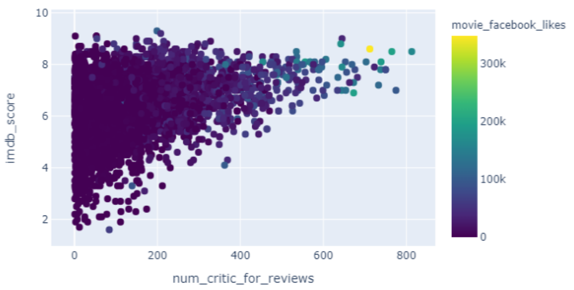

Box plot of aspect_ratio
imdb_score
9
8
7
6
5
4
3
2
1.18
1.2
1.75
4.0
2.76
2.24
2.4
16.0
1.44
2.2
1.33
1.66
1.37
1.5
2.0
2.39
2.55
1.85
2.35
1.77
1.78
1.89
aspect_ratio

The result of checking columns that has the number of unique values < 100 (except target column imdb_score):

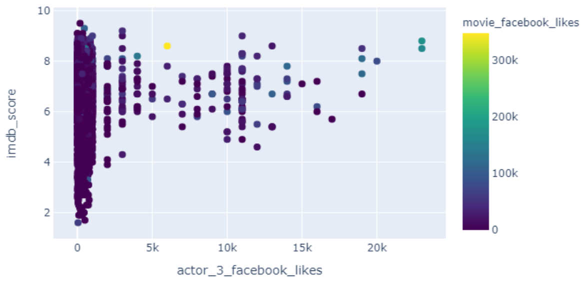
- Color column has 2 unique values. Most of movies are color movies (95%). Black and white accounts for 4 % and a small portion of Nan values. Black movies have higher median imdb score and most of them are > 5 while many color movies is rated below 5 (bad movie).
- Content_rating column has 18 types: R (41%), PG-13(29%), PG (14%), missing value (6%), and other values account a small proportion. The median rating is not different significantly between each type.
- Country column has 65 different values. Most movies come from USA and UK (total 84% for both), including a large amount of bad movies. The median imdb scores for these two countries are not the highest. Countries such as Portland, Iran, Sweden and Brazil, produced a small number of movies with higher median imdb scores.
- Language column has 47 different languages. Most movies are English (93%) because US and Uk are major countries for movie making. Many english movies has rating <5
- facenumber_in_poster has 19 different number. Most movies have less than 3 human faces. Median imbd score is also higher for number of faces < 3.
- aspect_ratio: most of them are 2.35 and 1.85, missing values (6.5%). The median rating is not different significantly between aspect_ratio.
- tilte_year: more than 90% of number movie was produced after 1995 when cinema industry thrived, mising values accounts 2.1%. Along with the boom of movie industry, after 2000 there are many movies with low imdb score.

- Check columns with unique variable > 100.
  - Scatter plots:

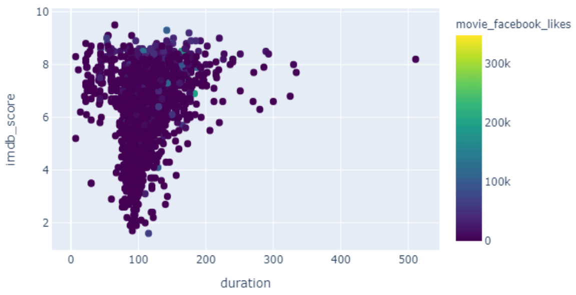
Scatter plot between Imdb_score and num_critic_for_reviews



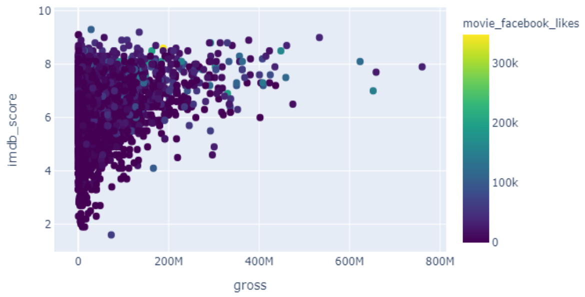
Scatter plot between Imdb_score and actor_3_facebook_likes



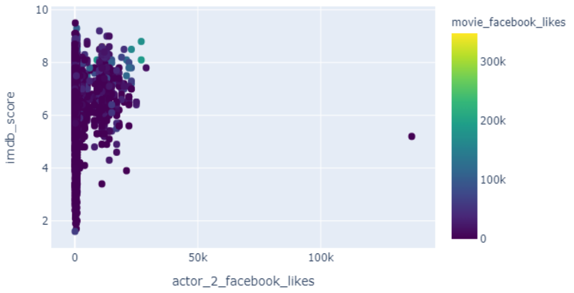
Scatter plot between Imdb_score and duration



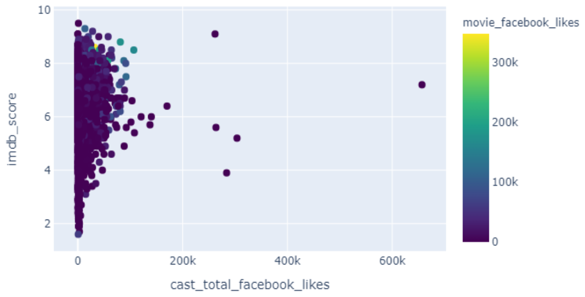
Scatter plot between Imdb_score and gross



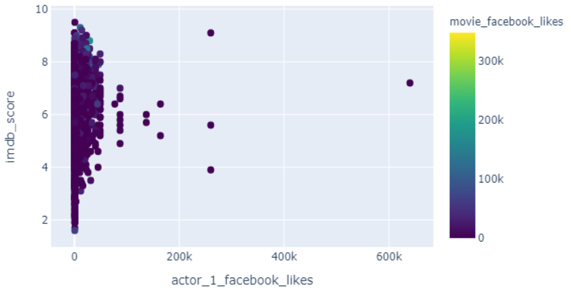
Scatter plot between Imdb_score and actor_2_facebook_likes



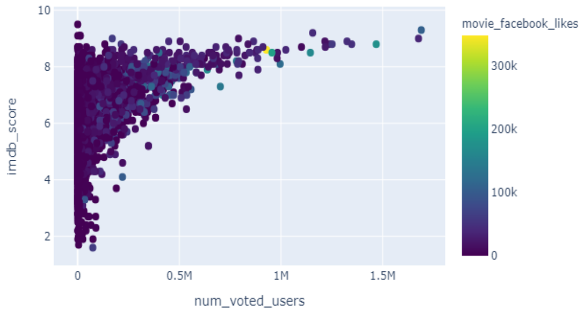
Scatter plot between Imdb_score and cast_total_facebook_likes



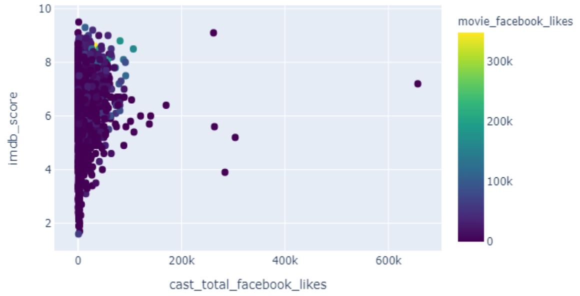
Scatter plot between Imdb_score and actor_1_facebook_likes



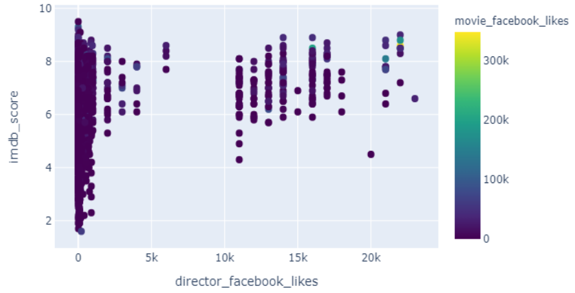
Scatter plot between Imdb_score and num_voted_users



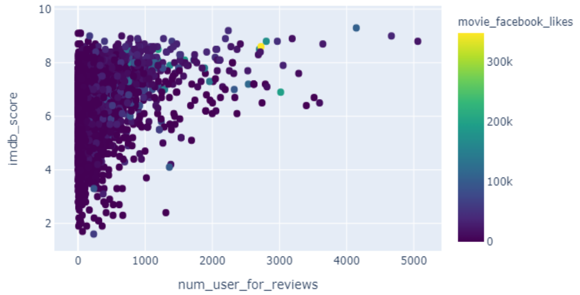
Scatter plot between Imdb_score and cast_total_facebook_likes



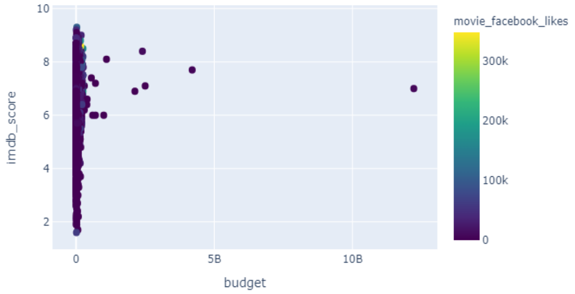
Scatter plot between Imdb_score and director_facebook_likes



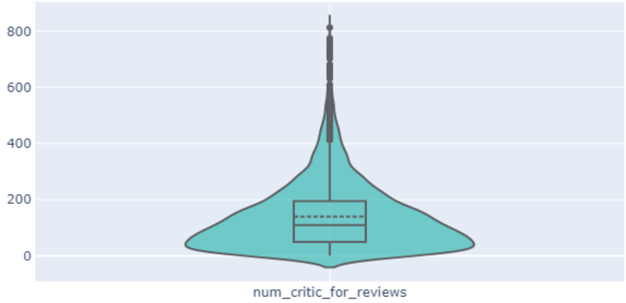
Scatter plot between Imdb_score and num_user_for_reviews

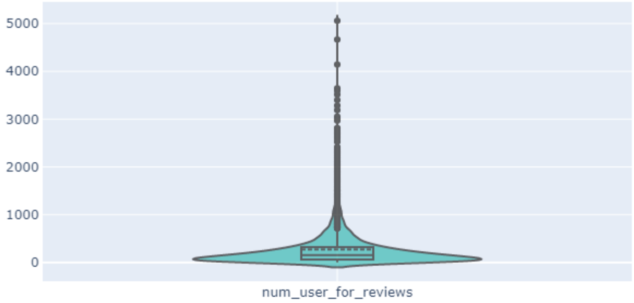


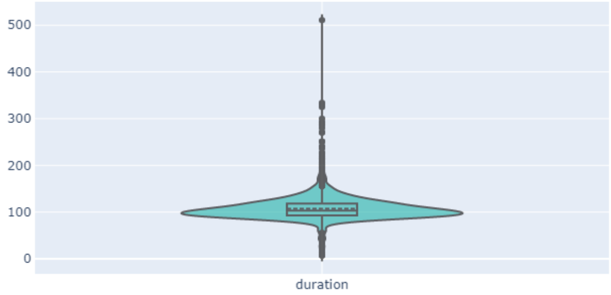
Scatter plot between Imdb_score and budget

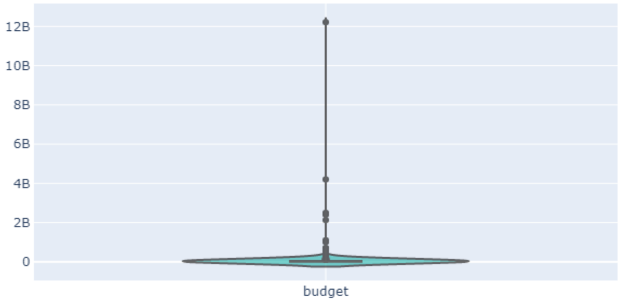


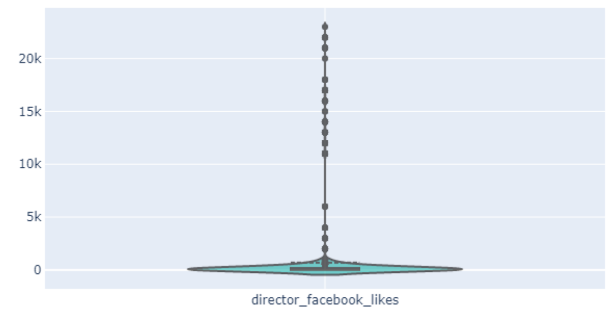
- Violin plots (box plot + density plot):

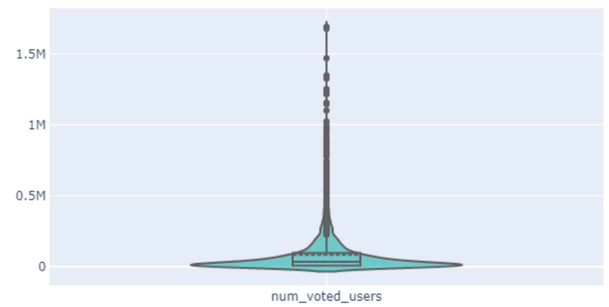


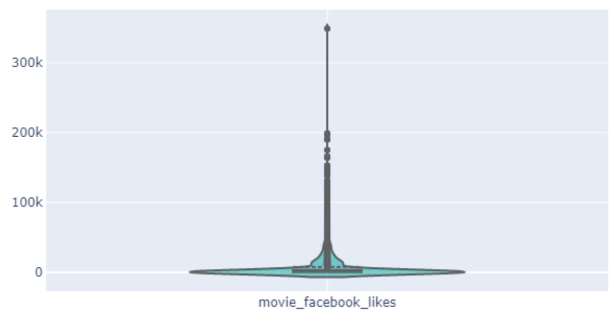


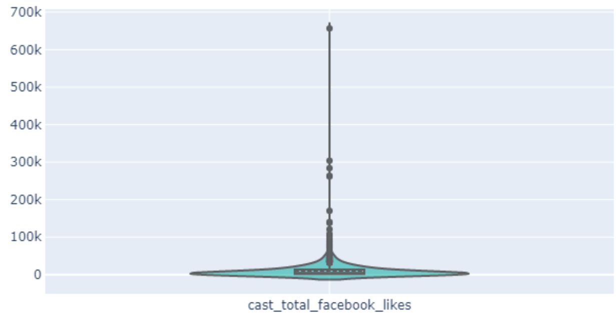


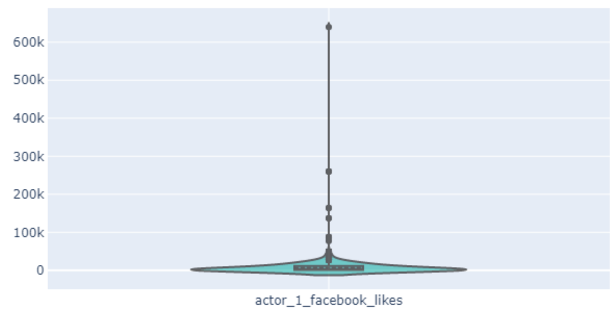


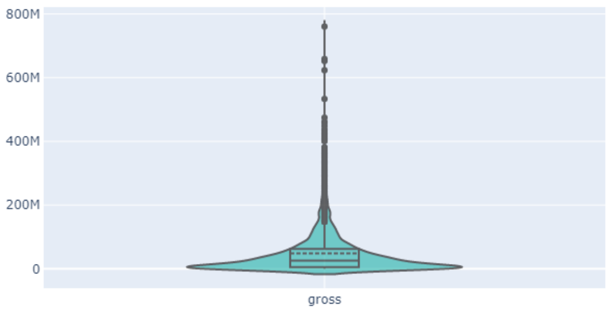

20k
15k
10k
5k
0
director_facebook_likes

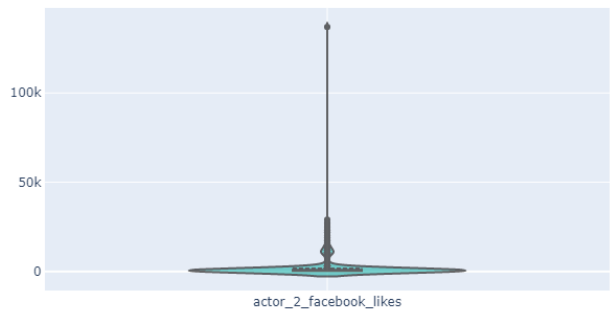

1.5M
1M
0.5M
0
num_voted_users

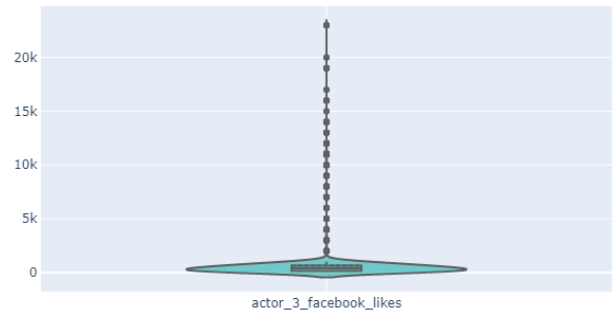

300k
200k
100k
0
movie_facebook_likes


700k
600k
500k
400k
300k
200k
100k
0
cast_total_facebook_likes


600k
500k
400k
300k
200k
100k
0
actor_1_facebook_likes


800M
600M
400M
200M
0
gross

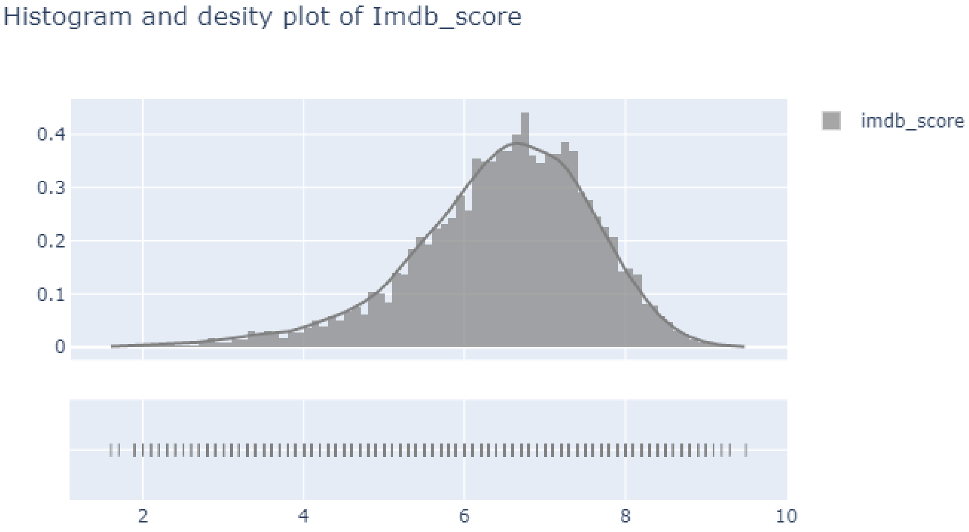

100k
50k
0
actor_2_facebook_likes


20k
15k
10k
5k
0
actor_3_facebook_likes

The result of checking columns that has the number of unique values > 100:

- Variable such as director_facebook_likes, movie_facebook_likes, num_critic_for_reviews, actor_3_facebook_likes,gross, num_voted_users, and num_user_for_reviews show a relationship with imdb rating, namely high values of these variables tends to go with high imdb rating, though still have movies that these variables are low but get high imdb rating.
- Other variables such as duration, actor_1_facebook_likes, actor_2_facebook_likes, cast_total_facebook_likes, cast_total_facebook_likes and budget seems not affect to imdb rating

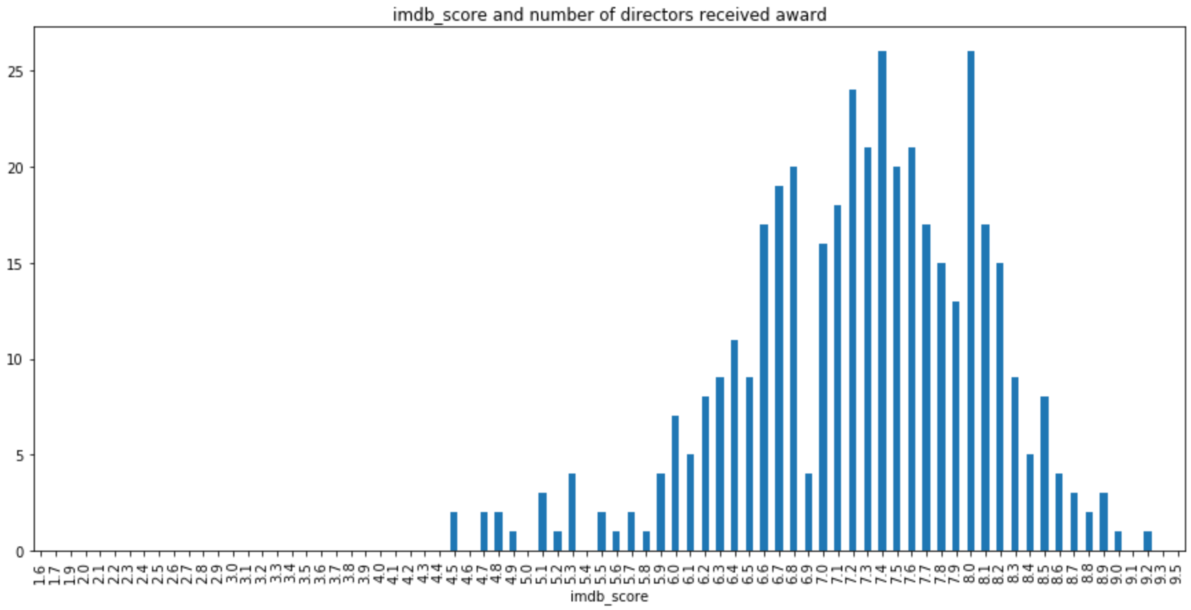
- Histogram chart of IMDB rating:

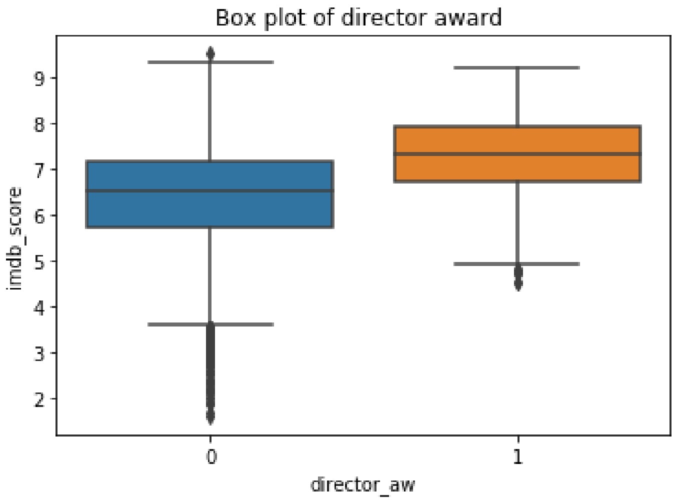


III. Adding a new variable

The greatness of a movie is highly affected by its director, so I will add a new column that can measure this property.

- Do web scraping from Wikipedia to get lists of winning Canne directors and winning Oscar directors.
- Merge 2 lists and remove duplicate names.
- Join back to the original data and create a new column with 1 means that director gets Canne or Oscar, and 0 is non





Generally, directors that received Canne/Oscar awards have higher Imdb score

IV. Transform categorical variables that will use in the model

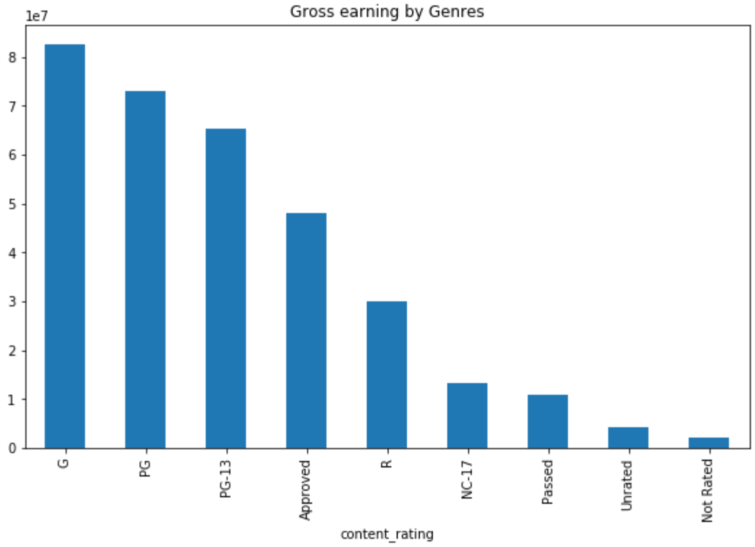
1. Content_rating columns

- M, GP, X are ratings used in the past. M and GP was replaced by PG, and X was replaced by NC-17.

We will group these content ratings into 7 main groups

- R: includes R, TV-MA
- PG-13: includes PG-13, TV-14
- G: includes G,TV-G, TV-Y, TV-Y7
- PG: includes PG, TV-PG, M, GP

- NC-17: includes NC-17, X
- nan
- Others



G, PG and PG-13 are genres that often bring higher profits

2. Language column

- Group into English, non English and nan.
- English if the movies come from countries that use English as an official language such as ['USA', 'UK', 'Canada', 'Australia', 'New Zealand']

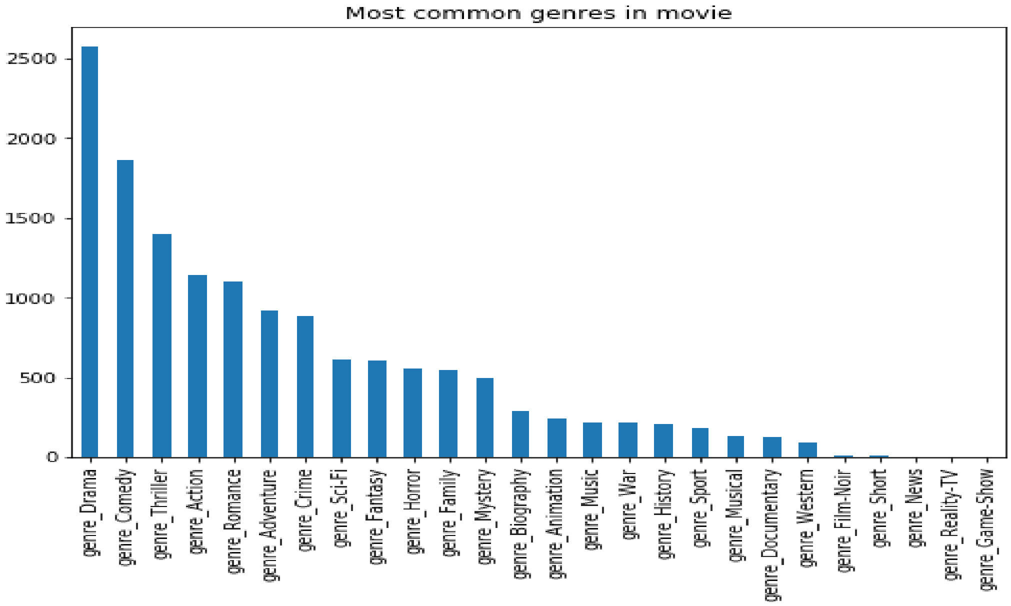
3. Country column

- Group into US, UK, France , nan, and others.

4. Genres column

- Each row of genre column contains different values, we will split these values, and get a unique genres list.

- Create corresponding dummy variables for these values.



Drama, comedy, thriller, and action are most common genres for movies.

IV. Remove unnecessary columns

- Our goal is to predict the Imdb rating at the time it's released, so the information that comes after such as gross earning won't be counted.
- cast_total_facebook_likes equals to sum of cast members facebook likes, so it will have high correlation with actor1, actor2, actor3 facebook likes and does not contribute additional information. We can drop it.
- Remove following variables: 'gross', 'movie_imdb_link', 'plot_keywords', 'director_name', 'actor_3_name', 'actor_2_name', 'actor_1_name', 'title_year', 'movie_title', 'genres', 'language', 'cast_total_facebook_likes'.

V. Handle missing data

For missing values

- There are some methods to handle missing values such as filling with mean for numerical variable, mode for categorical variabel, or knn imputer, though these methods often don’t give a good accuracy and for this case, missing values account for a small portion, so I will drop them.
- After drop all missing values, we still have 4136 rows (drop 17.9%).

VI. Create dummy variable for categorical variables.

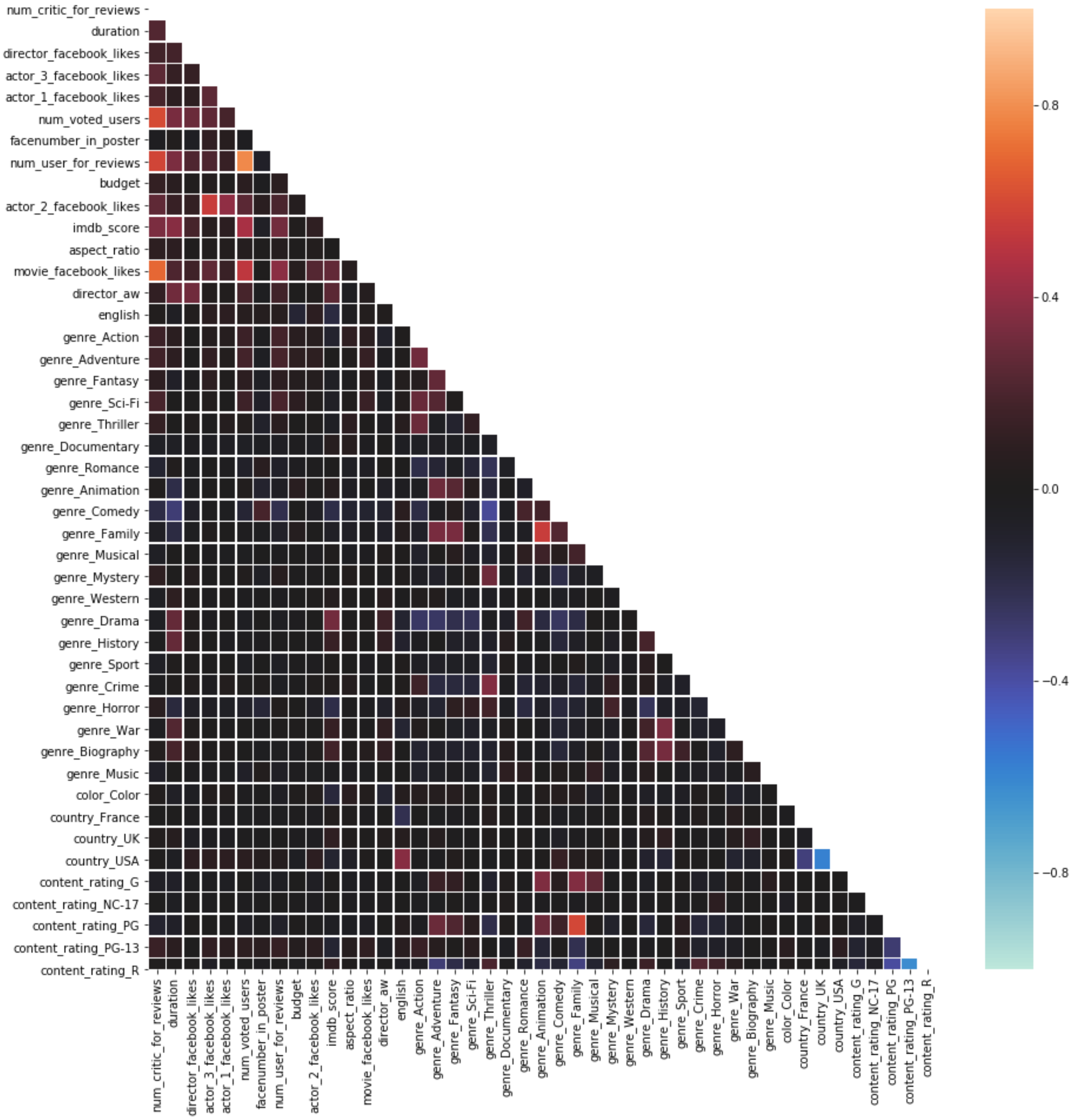
- Categorical variable will be transformed to dummy variables, and to avoid the high correlation between the dummy variables, I will drop one dummy variable of each category.

- Created dummy varibles: color_ Black and White, color_Color, country_France, country_UK, country_USA, country_others, content_rating_G, content_rating_NC-17, content_rating_PG, content_rating_PG-13,content_rating_R, content_rating_others
- Remove: color_ Black and White, country_others, content_rating_others

VII. Analyzing features

1. Correlation map

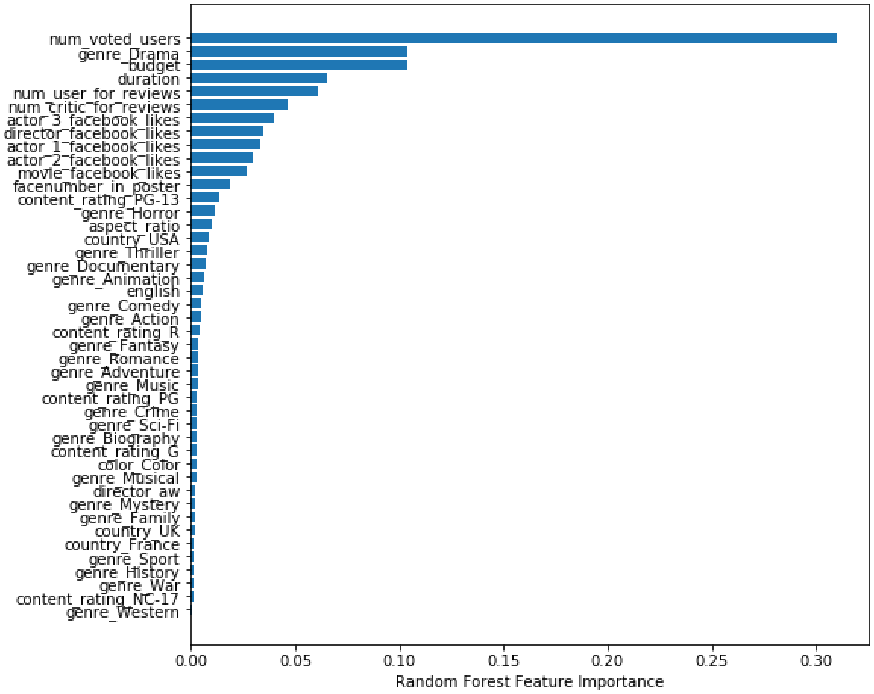
- Most variables don’t have high correlation

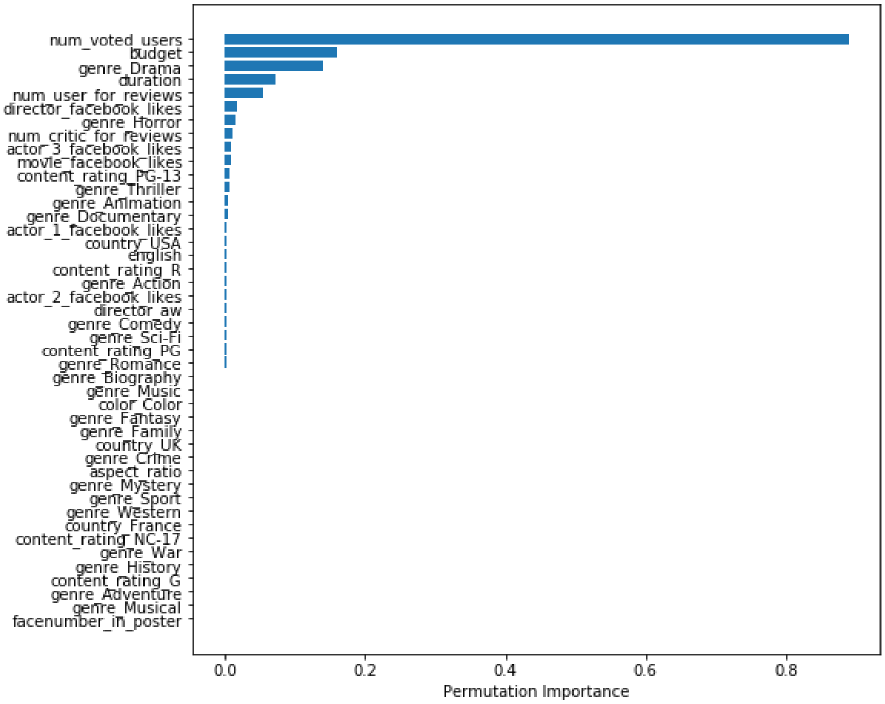


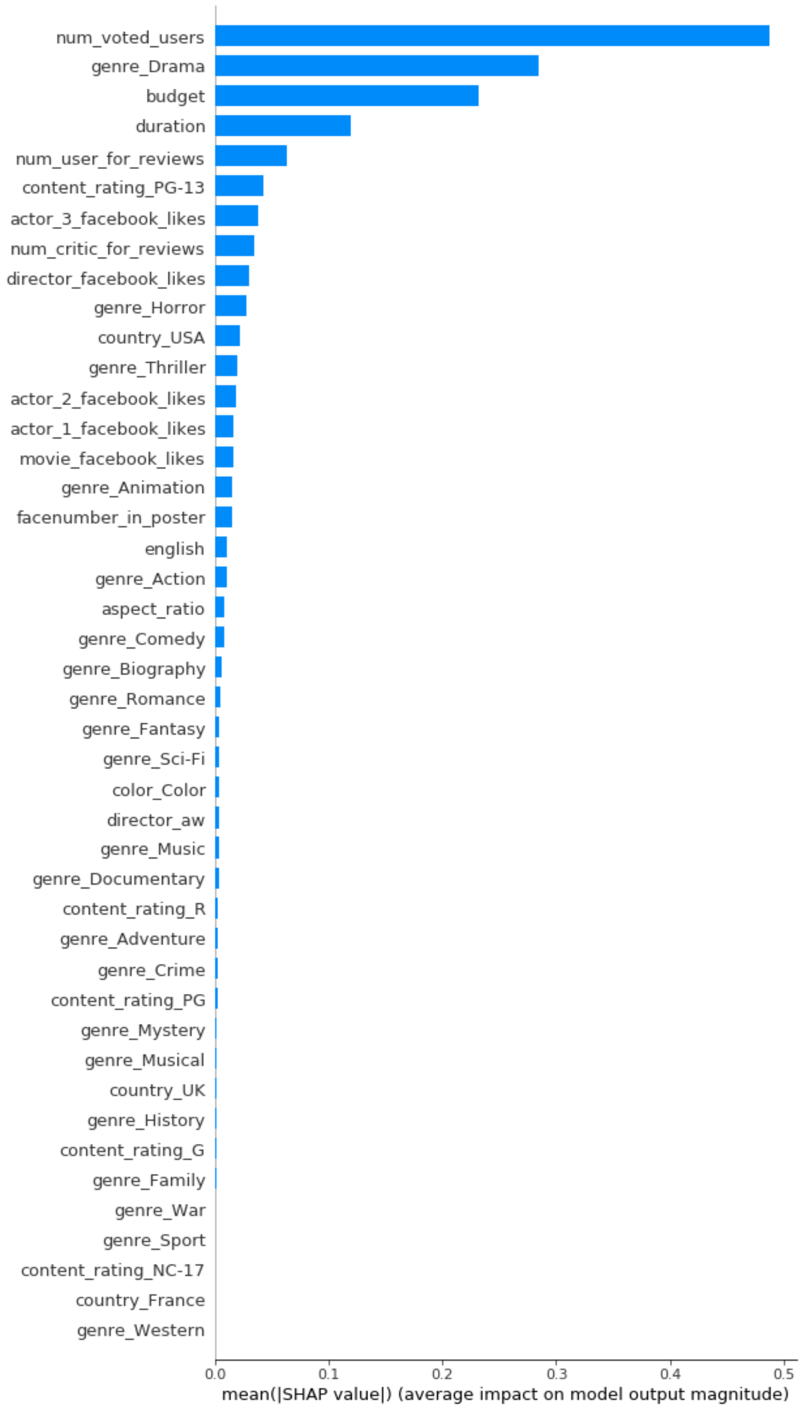
2. Feature importance

We will apply 3 ways to measure the important of features and compare the results:

- Random forest importance
- Permutation importance
- SHAP importance.







Num_voted_users, genres_drama, budget, duration, num_user_for_reviews are the most importance variables.

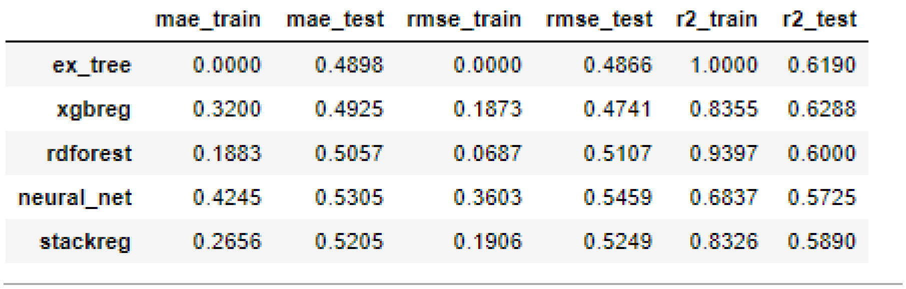
Take 25 variables that has highest important score from each method, combine and get their unique values. These variables are input in the models.

Selected variables: actor_2_facebook_likes, facenumber_in_poster, actor_1_facebook_likes, genre_Drama, country_USA,duration, genre_Thriller, num_critic_for_reviews, genre_Action, director_aw, num_user_for_reviews, genre_Animation, content_rating_PG, english, content_rating_PG-13, genre_Comedy, genre_SciFi, genre_Horror, budget, content_rating_R, movie_facebook_likes, genre_Documentary, actor_3_facebook_likes, aspect_ratio, director_facebook_likes, genre_Romance, num_voted_users.

VIII. Models

- Test data = 30% of dataset.
- Run regression to predict the Imdb score, using extra tree, xgboost, random forest, neural network and stacking regressor models.
- Train data will be divided into 5 k-folds to do cross validation. Data for xgboost and neural network will be standardized before training.
- Use GridSearchCV to select best parameters and then combine all the best models to run essemble stacking model.

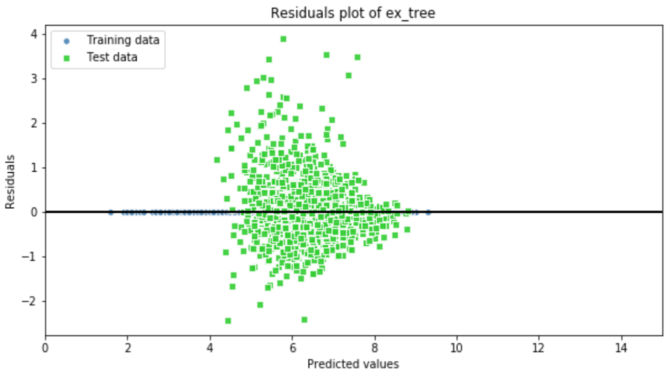
IX. Models evaluations and results

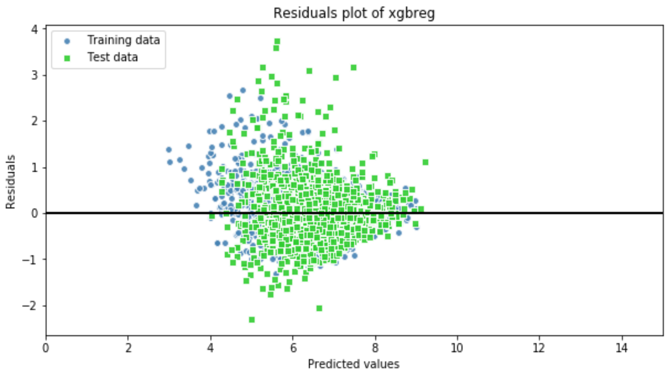


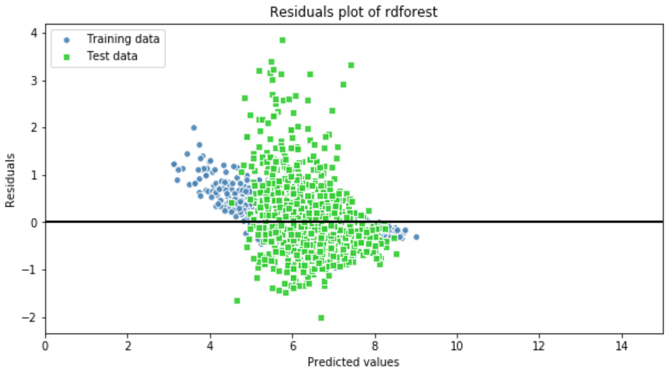
	mae_train	mae_test	rmse_train	rmse_test	r2_train	r2_test
ex_tree	0.0000	0.4898	0.0000	0.4866	1.0000	0.6190
xgbreg	0.3200	0.4925	0.1873	0.4741	0.8355	0.6288
rdforest	0.1883	0.5057	0.0687	0.5107	0.9397	0.6000
neural_net	0.4245	0.5305	0.3603	0.5459	0.6837	0.5725
stackreg	0.2656	0.5205	0.1906	0.5249	0.8326	0.5890

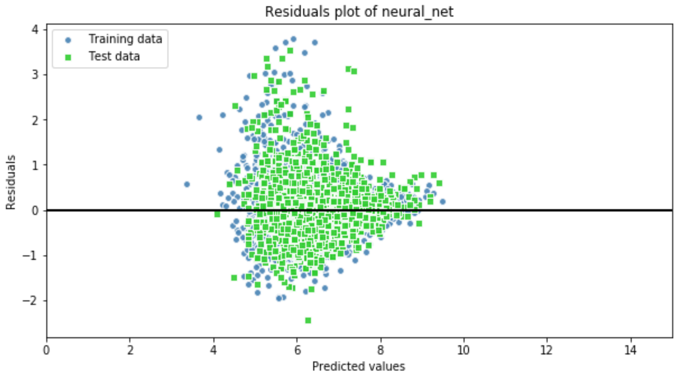
- Compare all metrics, the best model is Xgboost with the lowest MEA and RMSE value, and highest R square. For this model, all input variables can explain for 62% of the change in Imbd score.
- Extra tree has R_square_train = 1, and mea_train = 0 indicating it's a perfect fit, which we should suspect the result. The model show overfitting problem.

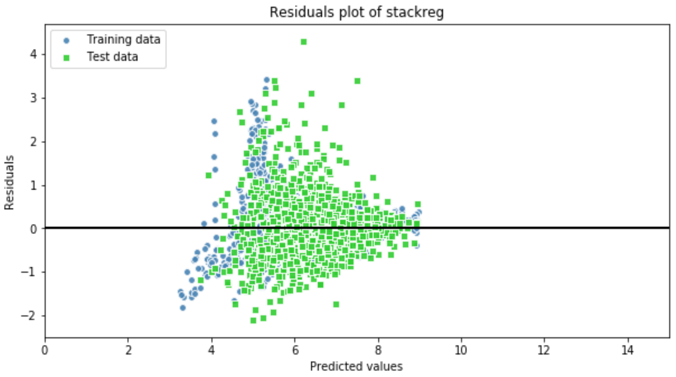
- Residuals plots





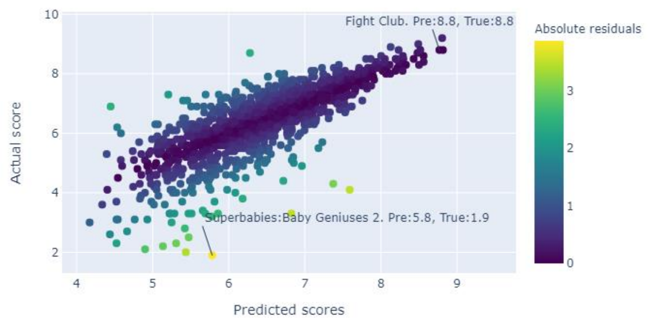


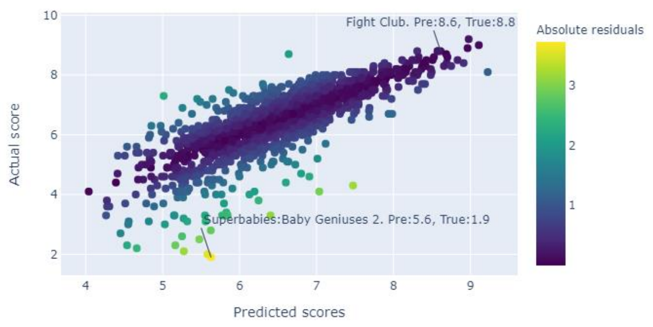


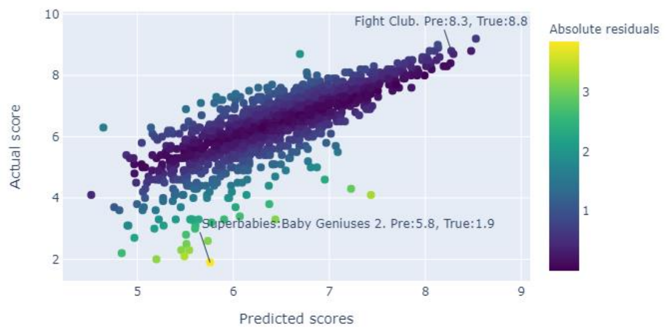


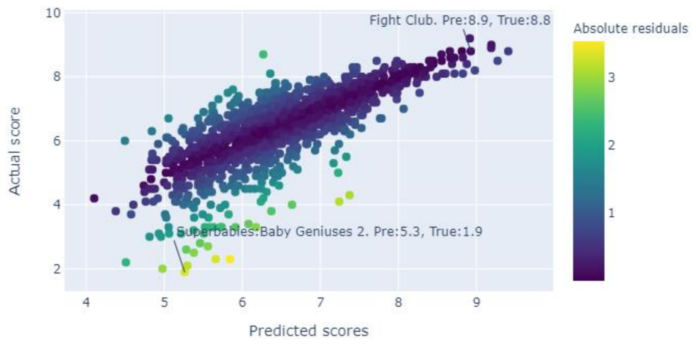
Overall, residuals of xgboost model has better shape.

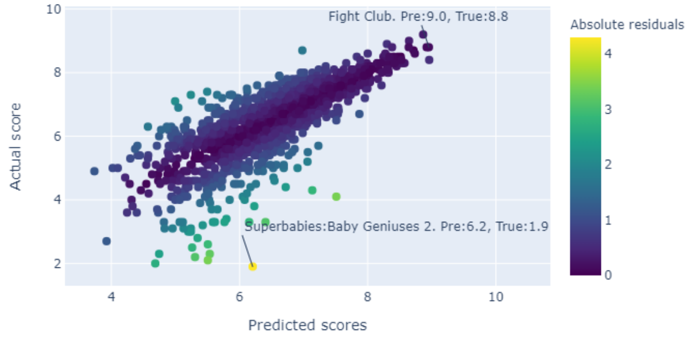
- Predicted and true Imdb score of each model


Imdb scores of ex_tree.MAE:0.49.RMSE:0.487. R2:0.619

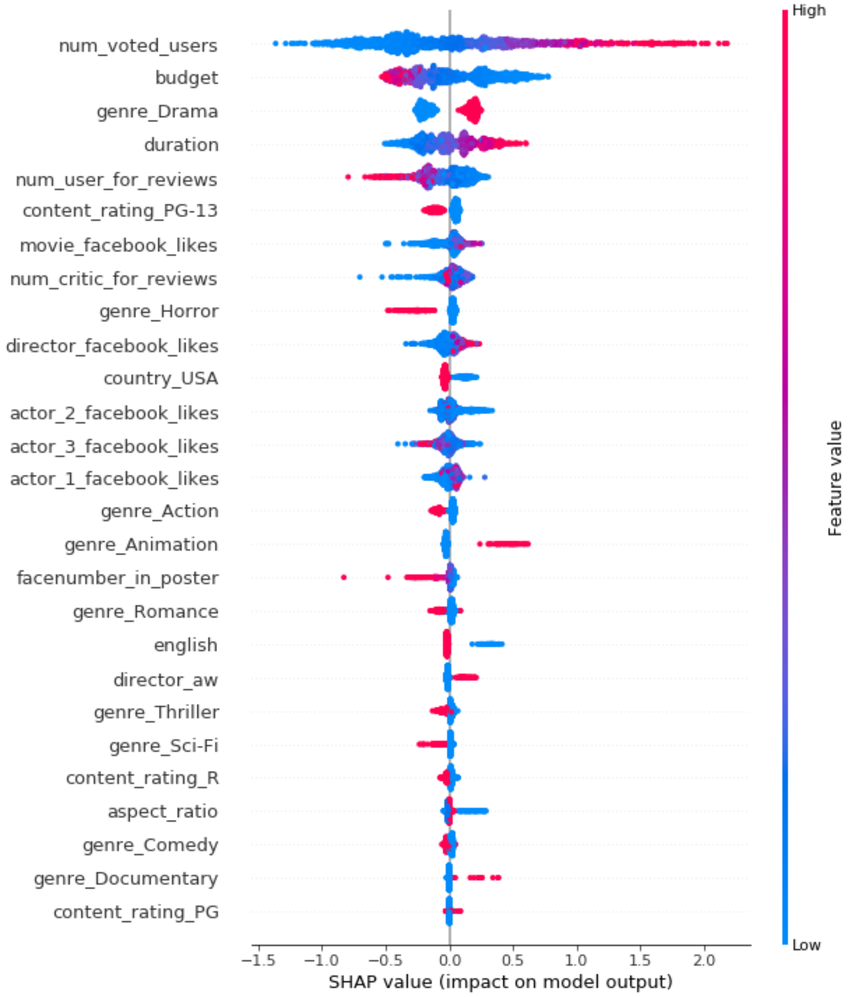

Imdb scores of xgbreg.MAE:0.492.RMSE:0.474. R2:0.629


Imdb scores of rdforest.MAE:0.506.RMSE:0.511. R2:0.6


Imdb scores of neural_net.MAE:0.531.RMSE:0.546. R2:0.573


Imdb scores of stackreg.MAE:0.521.RMSE:0.525. R2:0.589

In general, the prediction and true value forms a 45-degree line, which means their predicted and actual values are close. Movies such as Superbabies: Baby Geniuses 2 doesn't have a good prediction, but movie such as Fight Club performs well in predicting the score. We will check how input variables affect to the output of Xgboost model.



The order of y_axis indicates the importance of the features in the model, and how output changes when these feature changes its values:

- High num_voted_users will increase the imdb score. The same with long duration.
- High budget doesn’t mean high th imdb score.
- If a movie has drama genres, the imdb score will increase, while horror genres will decrease imdb score.
- High num_user_for_reviews will lower the imdb score, which we should consider carefully.
- If director receive oscar/canne award, imdb score is increased.
- Content_rating PG-13 and R will decrease imdb score
- More facenumber_in_poster will decrease the imdb score.

Conclusion:

- Data processing is heavy on handling categorical variables with many different values.

- We can combine 3 methods of feature importance to select variables because each method has its own merits and drawback such as random forest tends to prefer numerical features and categorical features with high cardinality, and in the case of correlated features it can select one of the feature and neglect the importance of the second one which can draw to a wrong conclusion.
- Xgboost is fast in training model and often gives a good predicting results. Random forest and neural network are also strong models to use in this case.
- Because the data has mix numerical and many binary dummy variables, so PCA is recommend not to use in this case to reduce the number of variables. PCA is desinged for continuous variables. It tries to break down the variance structure of a group of variables, and binary variables have no variance structure.