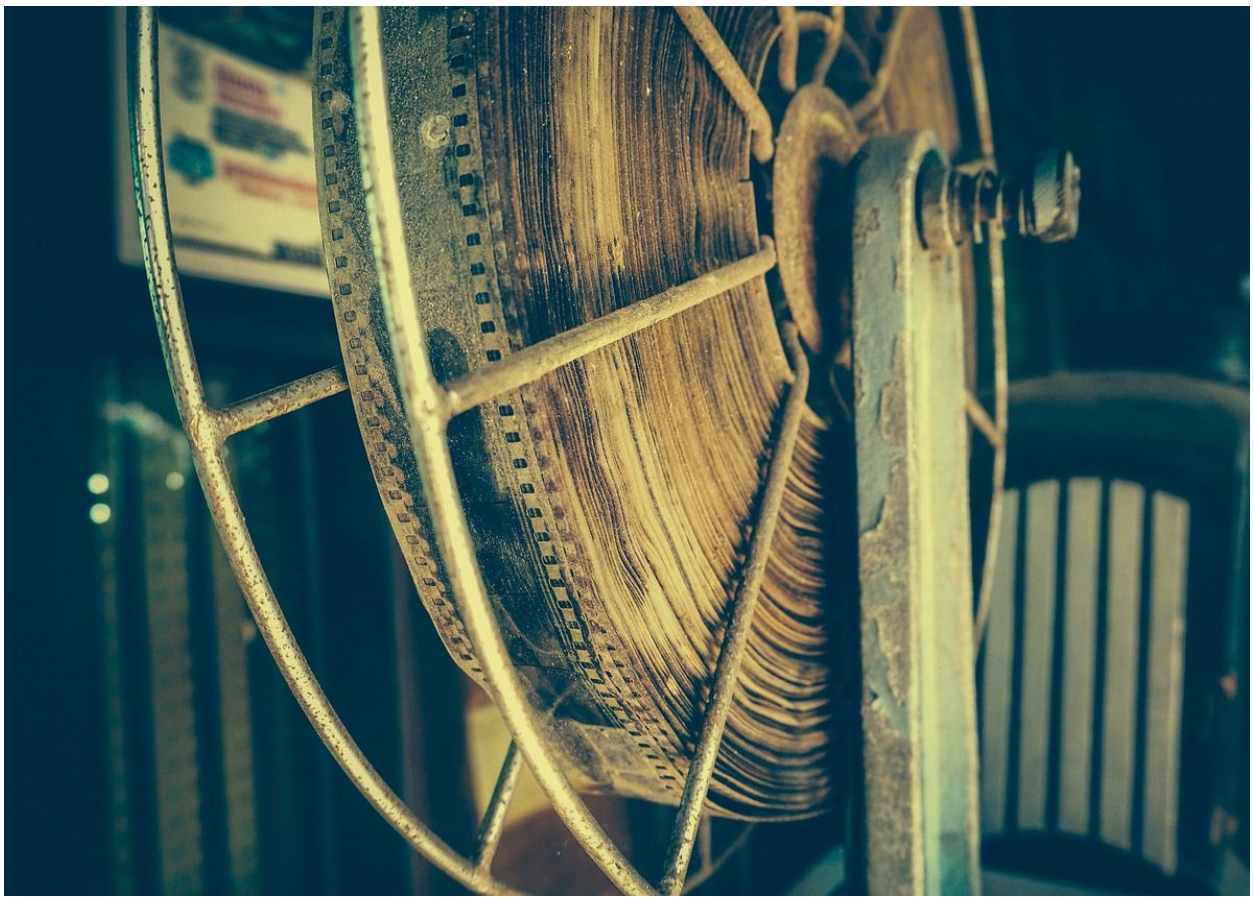
Business Analytics and Data Mining

Final Project

Factors Behind Successful Movies



Yun Mai
Gurpreet Singh
Chirag Vithalani
CUNY (City University of New York)
New York, NY 10017
Advisor / Guide: Marcus Ellis

Introduction

The motion picture industry is growing at a rapid growth rate, likely due to the acceleration of online and mobile distribution, lower admission prices, and government policy initiatives. This industry is also rich in data, thus making it extremely exciting for statisticians. The movie industry, which used to rely on traditional conventional wisdom and simple rules of thumb to predict box office outcomes, is slowly seeking new "analytical" approaches.

More and more analytical models will play a greater role in the motion picture industry by contributing towards superior marketing strategies that better predict the overall success of each movie.

Problem Statement

What makes movies good or bad? Is it our emotional response towards them? Is it the critical reviews or the scores? Is it the association of popular directors or actors? Is it the amount they gross at the box office? What is it really that describes their success or failure?

Factors Behind Successful Movies, as the name suggests, is an endeavor towards performing exploratory data analysis on a movie related dataset. Our intentions behind this exploration are to simply study a dataset that could provide some insight into movies, its audiences and to an extent it's commerce.

We will predict gross amount generated by the movies in US dollars. Revenue generated by a movie is the indicator of a movie to be successful. The greater value of gross, greater will be the popularity of the movie. In the dataset movies, there are 28 variables. Variable gross is our response variable and remaining 27 variables are predictor variables.

Data Collection

This data set was found from Kaggle. It contained data from 5043 movies spread across 28 features scraped from www.imdb.com. Some missing values is collected from https://www.the-numbers.com, http://www.boxofficemojo.com and http://en.wikipedia.org/wiki.

Features found in the dataset

Below is a complete list of features found in the dataset

- movie_title: **Contains title of a movie**
- color: **Specifies whether a movie is black and white or color**
- num_critic_for_reviews: **Contains number of critic reviews per movie**
- movie_facebook_likes: **Contains number of facebook likes per movie**
- duration: **Contains duration of a movie in minutes**
- director_name: **Contains name of the director of a movie**
- director_facebook_likes: **Contains number of facebook likes for a director**
- actor_3_name: **Contains the name of the 3rd leading actor of a movie**
- actor_3_facebook_likes: **Contains number of facebook likes for actor 3**
- actor_2_name: **Contains name of 2nd leading actor of a movie**
- actor_2_facebook_likes: **Contains number of facebook likes for actor 2**
- actor_1_name: **Contains name of the actor in lead role**
- actor_1_facebook_likes: **Contains number of facebook likes for actor 1**
- gross: **Contains the amount a movie grossed in USD**
- genres: **Contains the sub-genres to which a movie belongs**
- num_voted_users: **Contains number of users votes for a movie**
- cast_total_facebook_likes: **Contains number of facebook likes for the entire cast of a movie**
- facenumber_in_poster: **Contains number of actor's faces on a movie poster**
- plot_keywords: **Contains key plot words associated with a movie**
- movie_imdb_link: **Contains the link to the imdb movie page**
- num_user_for_reviews: **Contains the number of user generated reviews per movie**
- language: **Contains the language of a movie**
- country: **Contains the name of the country in which a movie was made**
- content_rating: **Contains maturity rating of a movie**
- budget: **Contains the amount of money spent in production per movie (not always in USD)**
- title_year: **Contains the year in which a film was released**
- imdb_score: **Contains user generated rating per movie**
- aspect_ratio: **Contains the size of the aspect ratio of a movie**

- **Movie facebook like: Contains number of facebook likes for a movie**

**Only the local Box Office is collected. That is, only the gross in USA was collected for a USA movie, only the gross in Asian was collected for a Asian movie and only the gross in Europe was collected for a European movie.*

Summary of the dataset

Highlights of main variables

Table1. Tally of Main Variables

5043 movies
100 years
66 countries
48 languages
2399 directors
28 variables

Variable Structure

Table2. Type of Variables

Variable	Structure
color	character
director_name	character
num_critic_for_reviews	integer
duration	integer
director_facebook_likes	integer

actor_3_facebook_likes	integer
actor_2_name	character
actor_1_facebook_likes	integer
gross	integer
genres	character
actor_1_name	character
movie_title	character
num_voted_users	integer
cast_total_facebook_likes	integer
actor_3_name	character
facenumber_in_poster	integer
plot_keywords	character
movie_imdb_link	character
num_user_for_reviews	integer
language	character
country	character
content_rating	character
budget	integer
title_year	integer
actor_2_facebook_likes	integer
imdb_score	numeric
aspect_ratio	numeric
movie_facebook_likes	integer

Word cloud for genres

First we would like to see which kind of movies being made more in general. As we can see there are less number of family movies, less documentaries and very few history related movies. More and more movies are of genres action,comedy, romance or Thriller.

Figure 1. Genres of the movies in the dataset.

All variables summarized

	vars	n	mean	sd	median	trimmed
num_critic_for_reviews	1	3996	139.60	121.84	110.00	121.39
duration	2	4023	107.03	24.55	103.00	105.08
director_facebook_likes	3	3951	674.07	2781.40	49.00	104.49
actor_3_facebook_likes	4	4015	665.64	1718.68	374.00	389.91
actor_1_facebook_likes	5	4029	6463.80	11234.74	989.00	4317.17
gross	6	3326	48196210.35	68291369.23	25001536.00	33690799.03
num_voted_users	7	4035	83593.84	140941.22	33180.00	52385.99
cast_total_facebook_likes	8	4035	9668.58	15238.49	3104.00	6562.14
facenumber_in_poster	9	4024	1.37	2.05	1.00	0.97
num_user_for_reviews	10	4018	271.68	382.89	153.00	195.13
budget	11	3638	39627490.43	224695698.13	20000000.00	24851004.07
title_year	12	3947	2002.43	12.44	2005.00	2004.56
actor_2_facebook_likes	13	4024	1691.30	4195.39	595.00	646.69
imdb_score	14	4035	6.45	1.12	6.60	6.52
aspect_ratio	15	3771	2.21	1.35	2.35	2.11
movie_facebook_likes	16	4035	7474.67	19008.09	167.00	2747.38

	mad	min	max	range	skew	kurtosis
num_critic_for_reviews	99.33	1.00	8.130000e+02	8.120000e+02	1.55	3.02
duration	17.79	7.00	3.340000e+02	3.270000e+02	1.69	10.95
director_facebook_likes	72.65	0.00	2.300000e+04	2.300000e+04	5.27	27.67
actor_3_facebook_likes	372.13	0.00	2.300000e+04	2.300000e+04	6.94	55.23
actor_1_facebook_likes	1149.01	0.00	2.600000e+05	2.600000e+05	6.25	92.86
gross	33772694.70	162.00	6.586723e+08	6.586721e+08	2.95	12.35
num_voted_users	44053.98	5.00	1.689764e+06	1.689759e+06	4.13	25.76
cast_total_facebook_likes	3454.46	0.00	3.037170e+05	3.037170e+05	5.16	62.29
facenumber_in_poster	1.48	0.00	4.300000e+01	4.300000e+01	4.77	59.13
num_user_for_reviews	164.57	1.00	5.060000e+03	5.059000e+03	4.22	27.77
budget	23721600.00	218.00	1.221550e+10	1.221550e+10	46.07	2404.93
title_year	8.90	1920.00	2.016000e+03	9.600000e+01	-2.24	6.98
actor_2_facebook_likes	474.43	0.00	1.370000e+05	1.370000e+05	10.53	274.17
imdb_score	1.04	1.70	9.500000e+00	7.800000e+00	-0.72	0.92
aspect_ratio	0.06	1.18	1.600000e+01	1.482000e+01	9.63	95.65
movie_facebook_likes	247.59	0.00	1.990000e+05	1.990000e+05	4.43	25.96

	se
num_critic_for_reviews	1.93
duration	0.39
director_facebook_likes	44.25
actor_3_facebook_likes	27.12
actor_1_facebook_likes	177.00
gross	1184144.49
num_voted_users	2218.79
cast_total_facebook_likes	239.89
facenumber_in_poster	0.03
num_user_for_reviews	6.04
budget	3725318.51
title_year	0.20
actor_2_facebook_likes	66.14
imdb_score	0.02
aspect_ratio	0.02
movie_facebook_likes	299.24

Data curation and cleaning

1. Aspect_ratio: the max value is 16, which looks like an outlier. Actually, 16 is introduced when scraping the information from IMDB website or IMDB API where only the first number of aspect ratio 16: 9 was collected. And the similar mistake is made for aspect ratio value 4, which is in fact 4 : 3. So 16 and 4 were replaced by 1.78 and 1.33 respectively.

2. Some of the missing gross values were filled by searching the information from other websites including https://www.the-numbers.com, http://www.boxofficemojo.com and http://en.wikipedia.org/wiki.

3. The hyphen in the names of the *genres* of and *content_rating* were replaced with an underscore because hyphen will be problematic when the variable names are used in the formula expression in R.

Data Preparation

Missing Values

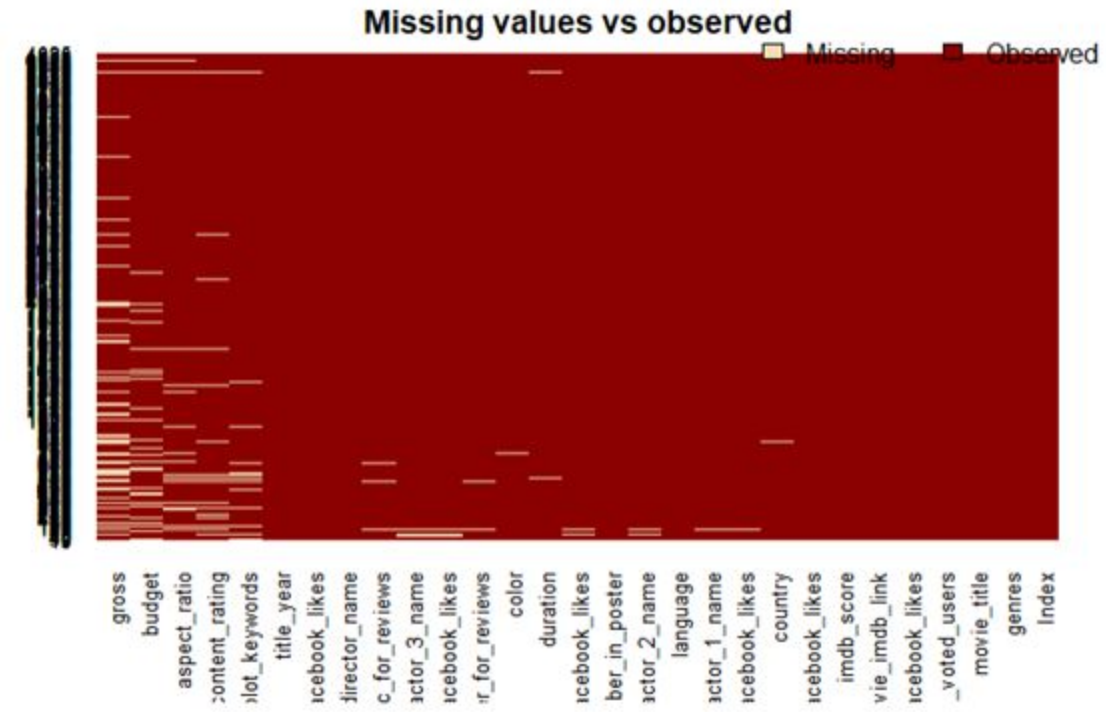


Figure 2. Missing data map.

Table 3. Missing Values

	feature_mode <fctr>	missing value <int>	missing value(%) <dbl>
gross	integer	788	15.65044687
budget	integer	456	9.05660377
aspect_ratio	numeric	309	6.13704071
director_facebook_likes	integer	101	2.00595829
num_critic_for_reviews	integer	49	0.97318769
actor_3_facebook_likes	integer	32	0.63555114
actor_2_facebook_likes	integer	22	0.43694141
num_user_for_reviews	integer	20	0.39721946
actor_1_facebook_likes	integer	16	0.31777557
cast_total_facebook_lik	integer	10	0.19860973
movie_facebook_likes	integer	9	0.17874876
duration	integer	1	0.01986097
movie_title	character	0	0.00000000
imdb_score	numeric	0	0.00000000
num_voted_users	integer	0	0.00000000
facenumber_in_poster	integer	0	0.00000000
content_rating	character	0	0.00000000
color	character	0	0.00000000
title_year	integer	0	0.00000000
language	character	0	0.00000000
country	character	0	0.00000000
director_name	character	0	0.00000000
actor_1_name	character	0	0.00000000
actor_2_name	character	0	0.00000000
actor_3_name	character	0	0.00000000
genres	character	0	0.00000000
plot_keywords	character	0	0.00000000
movie_imdb_link	character	0	0.00000000

From **Table 3** and **Figure 3,** we noticed that there are many missing values for gross, budget and aspect ratio. There are 1143 records have missing values and 405 records missed more than 5% data points. Same is depicted in below missing values graphics and histogram. Let's see missing data in more detail.

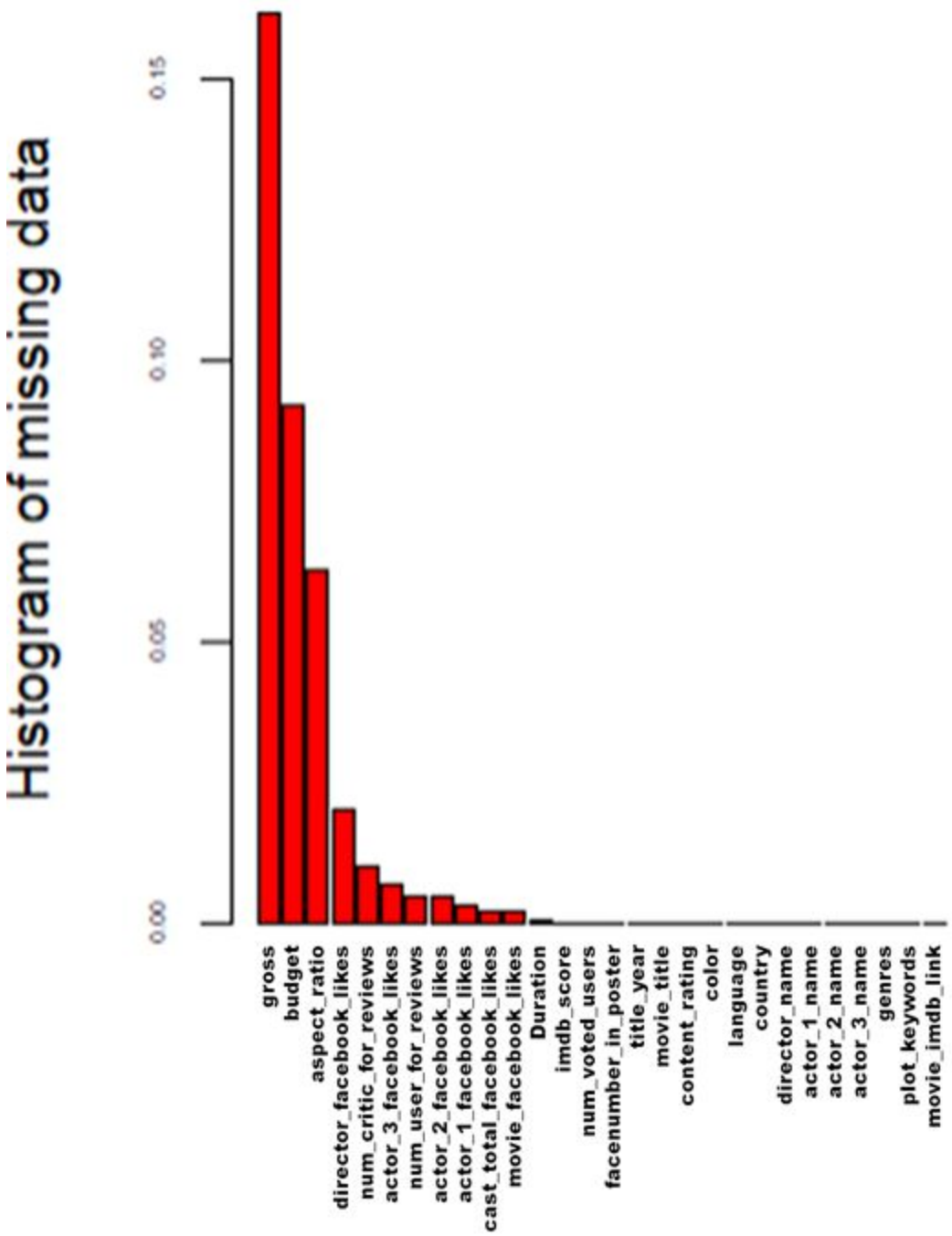


Figure 3. Missing data.

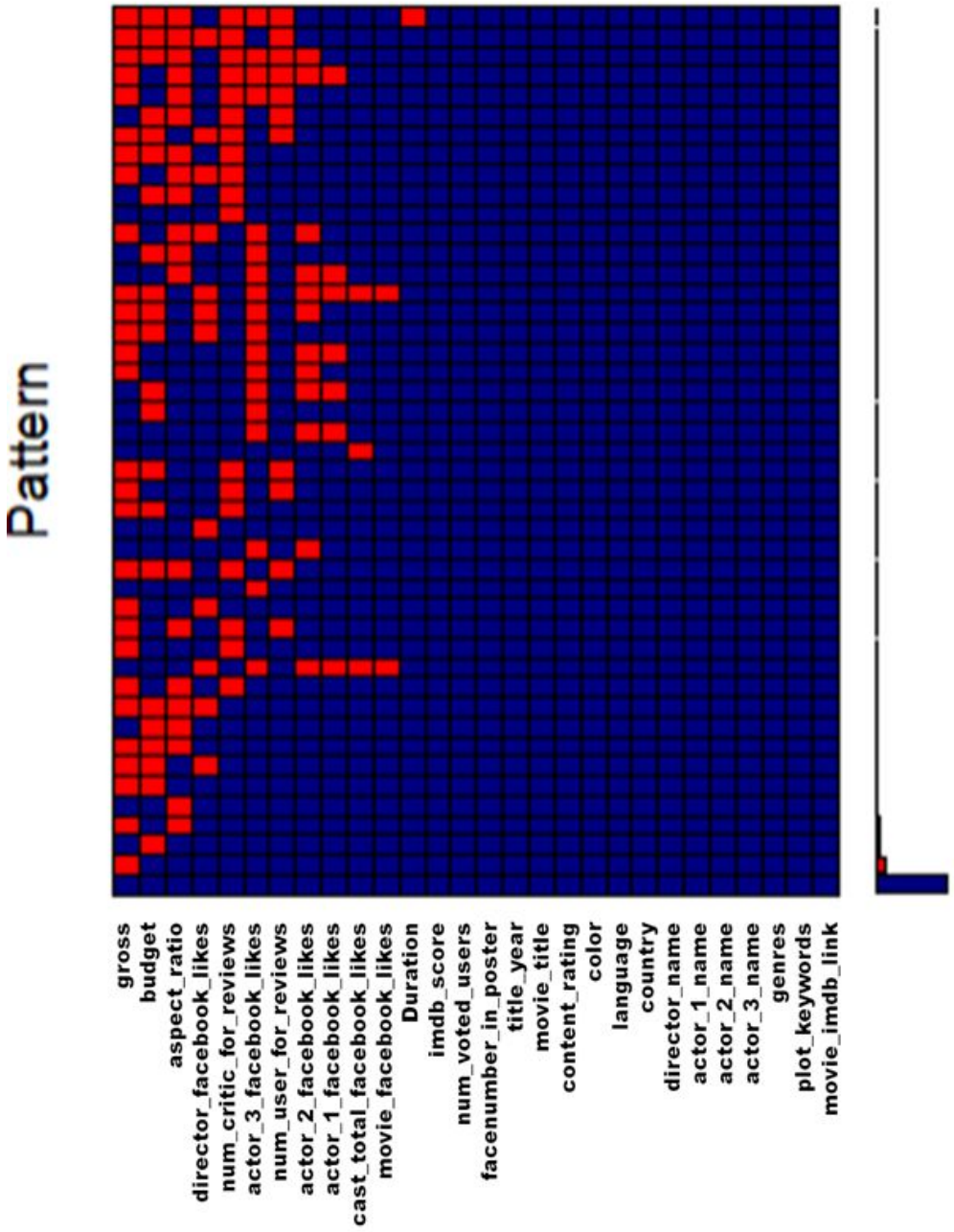


Figure 4. Missing data patterns.

For the categorical variables, if there is no value in director's name or NA, we will replace with the character 'NA'.

Variable ***gross*** has has the highest level of missing rate, 15.65%, in the dataset. Removing the variable from analysis might be a better choice. The heat-map (**Figure 4**) above shows

that the gross missed in many in almost all of the patterns. Multivariate imputation with chained equations will be done for the missing values. But because “gross” is our dependent variable, Missing Imputation and Deletion (MID) will be applied to the dataset. That is to say, the records with “gross” value missing will be deleted after imputation.

Imputation

Sample method was used to do multivariate imputation with chained equations. The distributions of the five imputed datasets (pink) are consistent to the original dataset as shown in the density graph below(**Figure 5**).

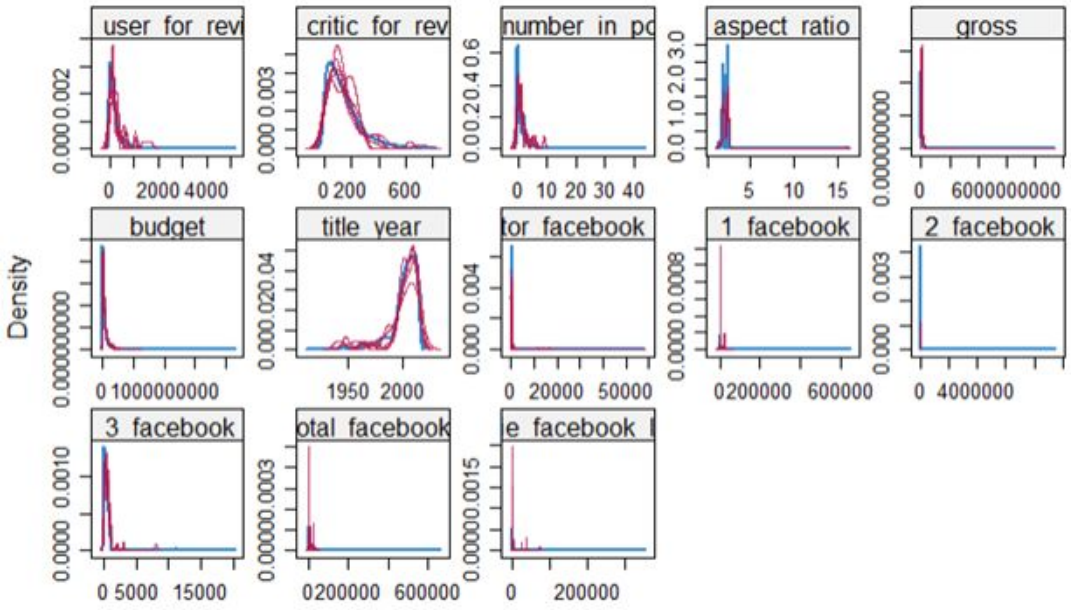


Figure 5. Imputation of the missing values by MICE.

After replacing NA with “NA” in the categorical variables and MID treatment for the numeric variables, we have completed dataset as shown below:

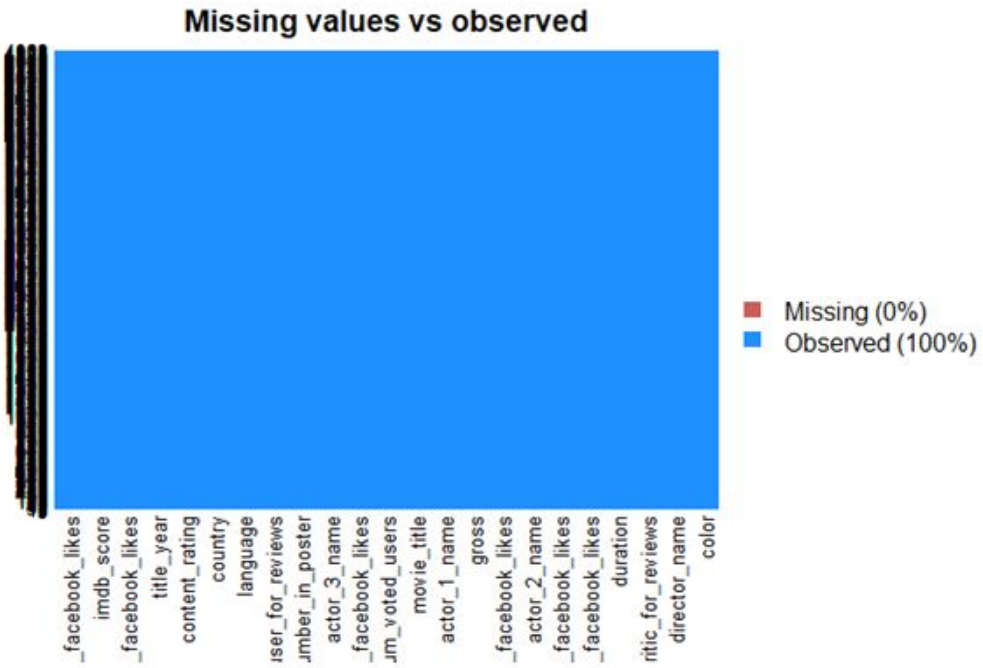


Figure 6. Missing map after imputation

Dependent Variable

Visualization of title Year vs. Gross

There are many outliers for title year. We have more data after year 1990 and the majority of data points are around the year of 2000 and later,which make sense which points there less movies in the early years. Also, it is interesting to notice that movies from early years tend to have higher scores(**Figure 7**).

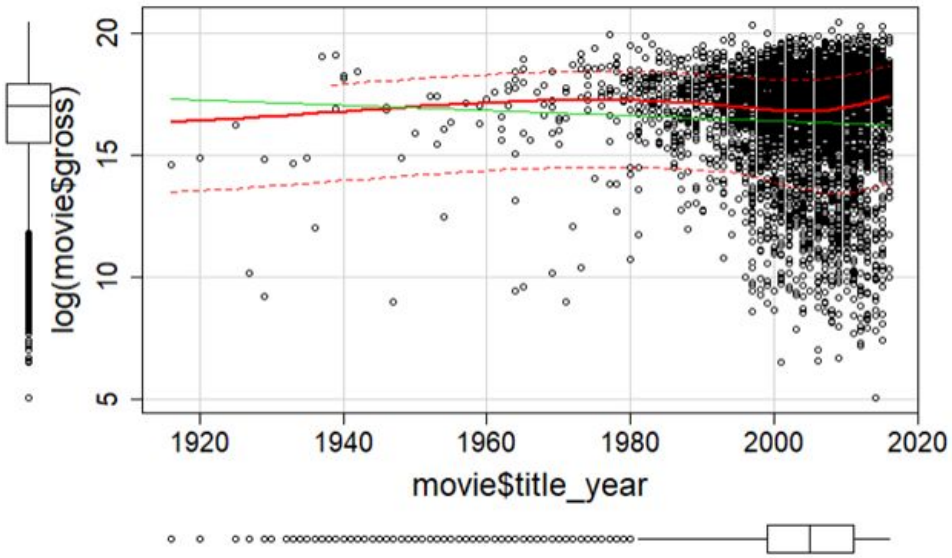


Figure 7. Movie gross in the past 100 years.

Visualization of the Distribution of *Gross*

gross is right-skewed. Many movies gross below 50 million and few movies have high gross more than 200 million (**Figure 8**).

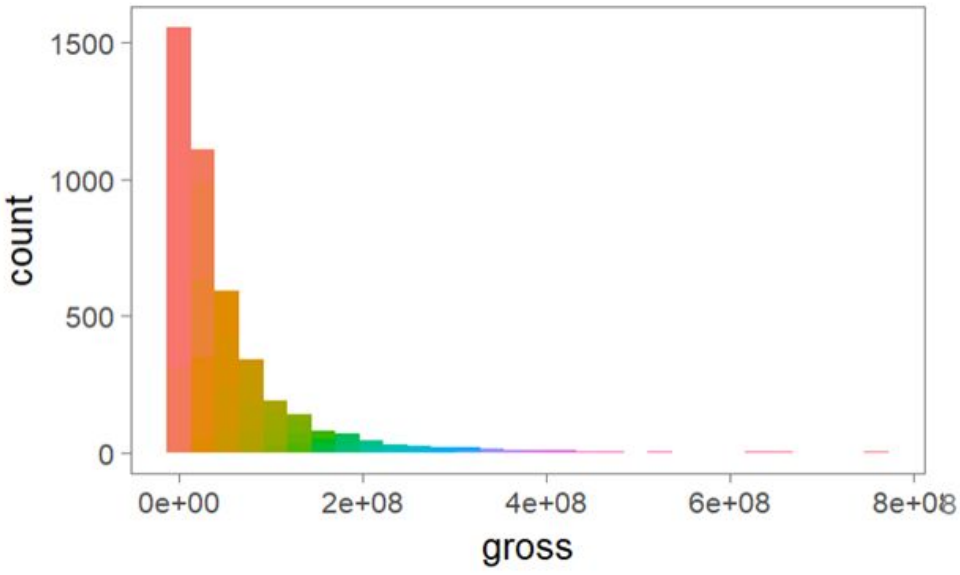


Figure 8. Distribution of movie gross.

Adding new variables

Apart from the readily available variables, some new variables were added: to the total number of movies and the average gross of previous movies of the actors and director involved. The newly added variables are: d***irector_ave_gross, actor_1_ave_gross, actor_2_ave_gross, actor_3_ave_gross, director_ave_sum, actor_1_ave_sum, actor_2_ave_sum, actor_3_ave_sum.*** The distribution of the newly generated variables is all right-skewed (Figure 9). It is not surprised because only a few very famous directors and actors were in those movies with the very high box office. Most of the directors and actors worked in the movies with box office less than 50 million.

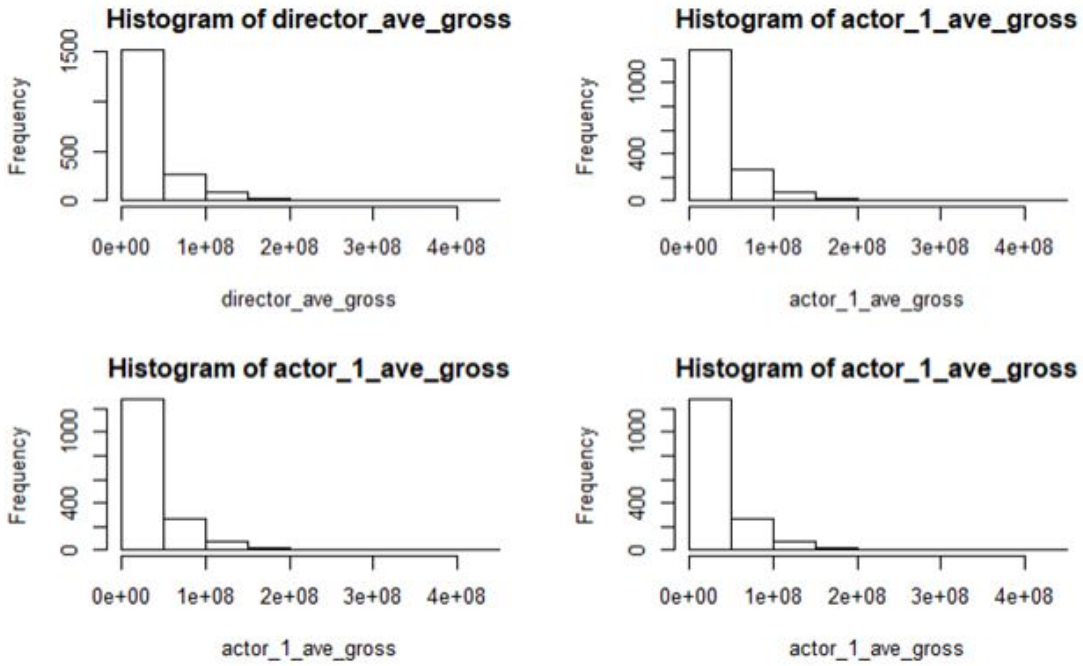


Figure 9. Distribution of the newly generated average gross of director and actors.

Exploring correlation

Blind guessing

We tend to think that abundant funding, famous producer, writers, director and popular actors will be the reasons make the success of a movie. So we first do a little blind guess to see whether *budget, imdb_score, director_facebook_likes, and actor_1_facebook_likes* have

positive correlation with *gross*. It looks like that *budget, director_facebook_likes and actor_1_facebook_likes* are correlated to *gross*.

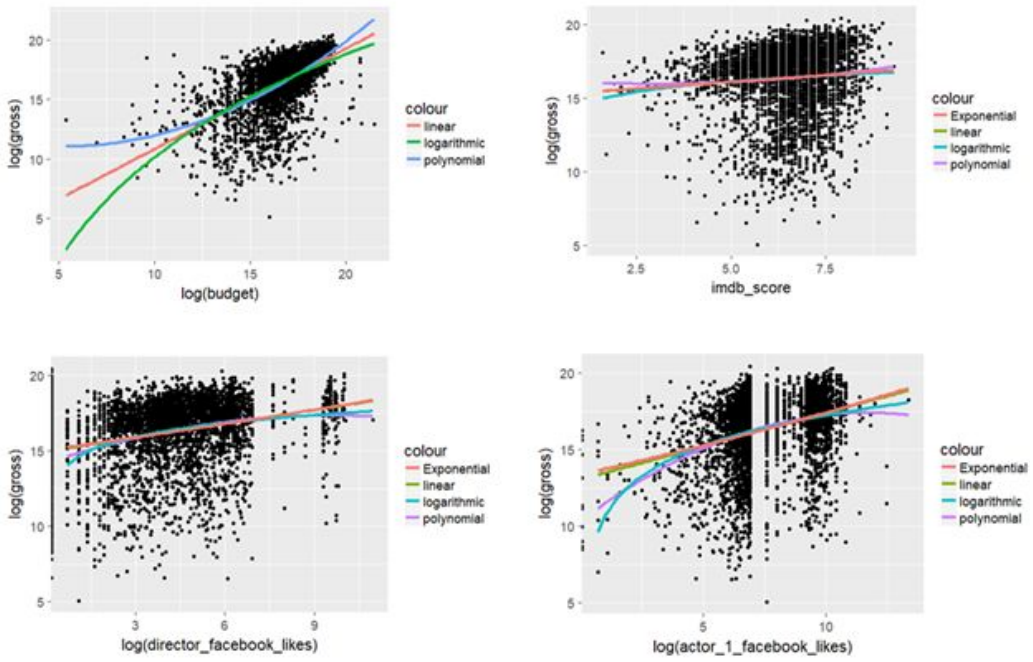
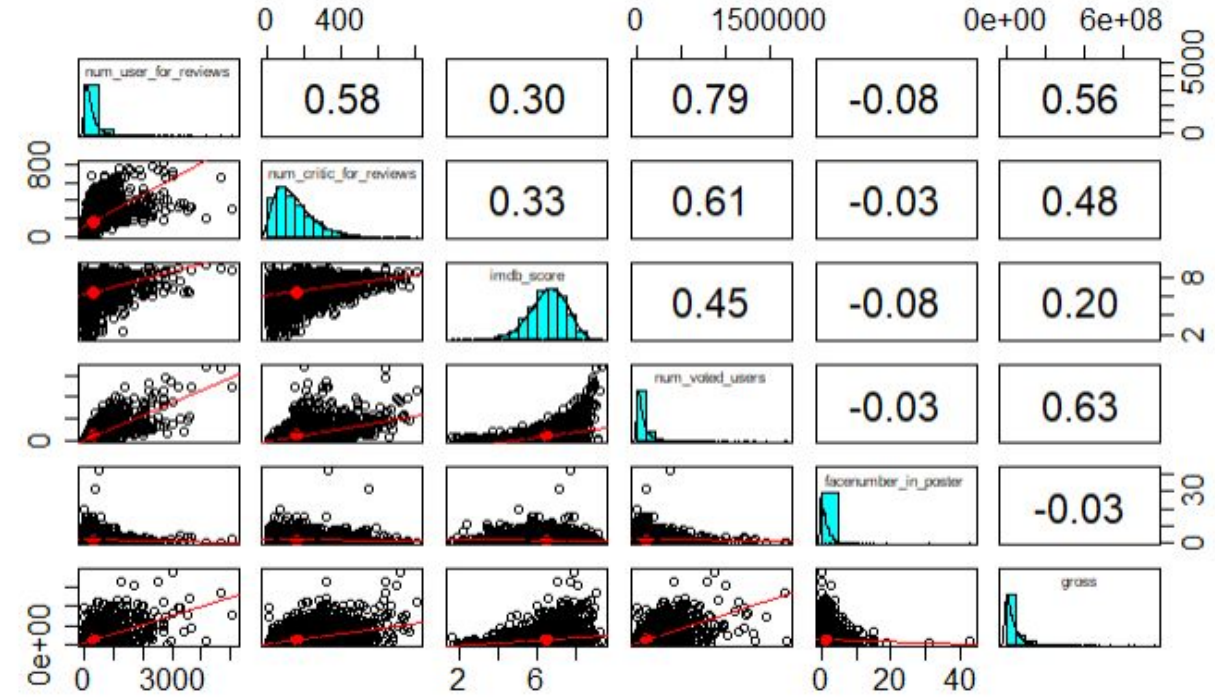
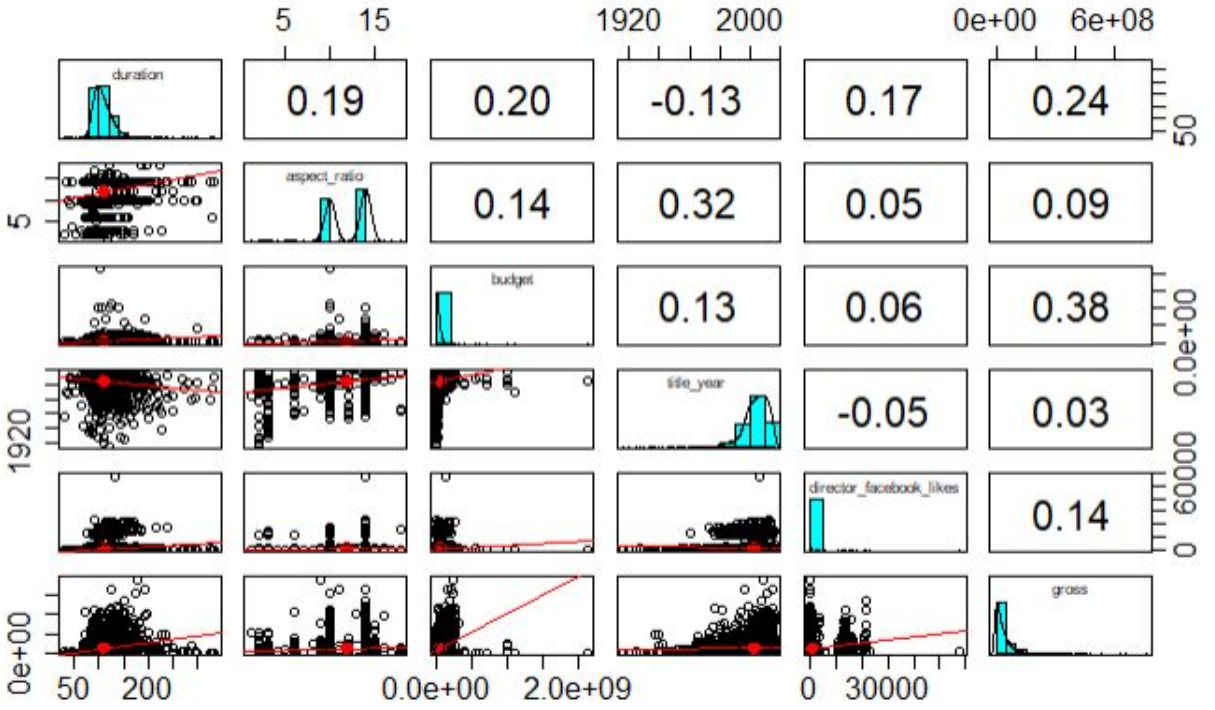


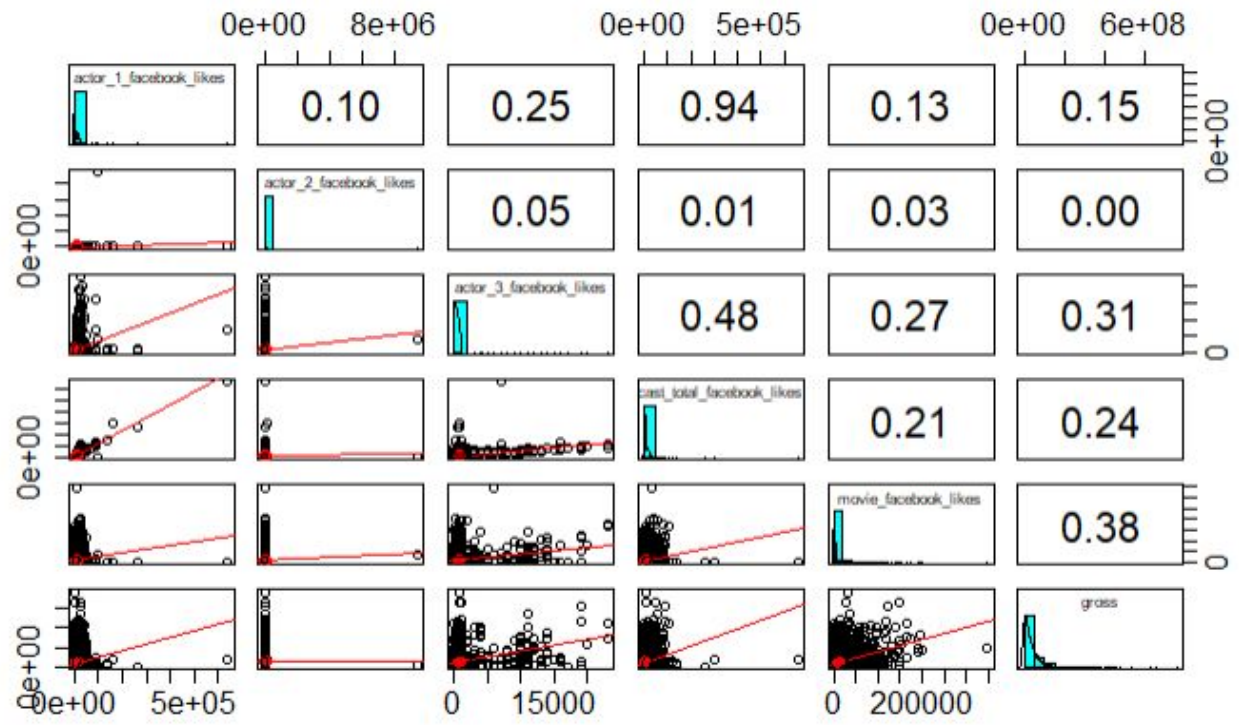
Figure 10. Explore the correlation between 4 independent variables budget, imdb_score, director_facebook_likes, actor_1_facebook_likes and dependent variable gross.

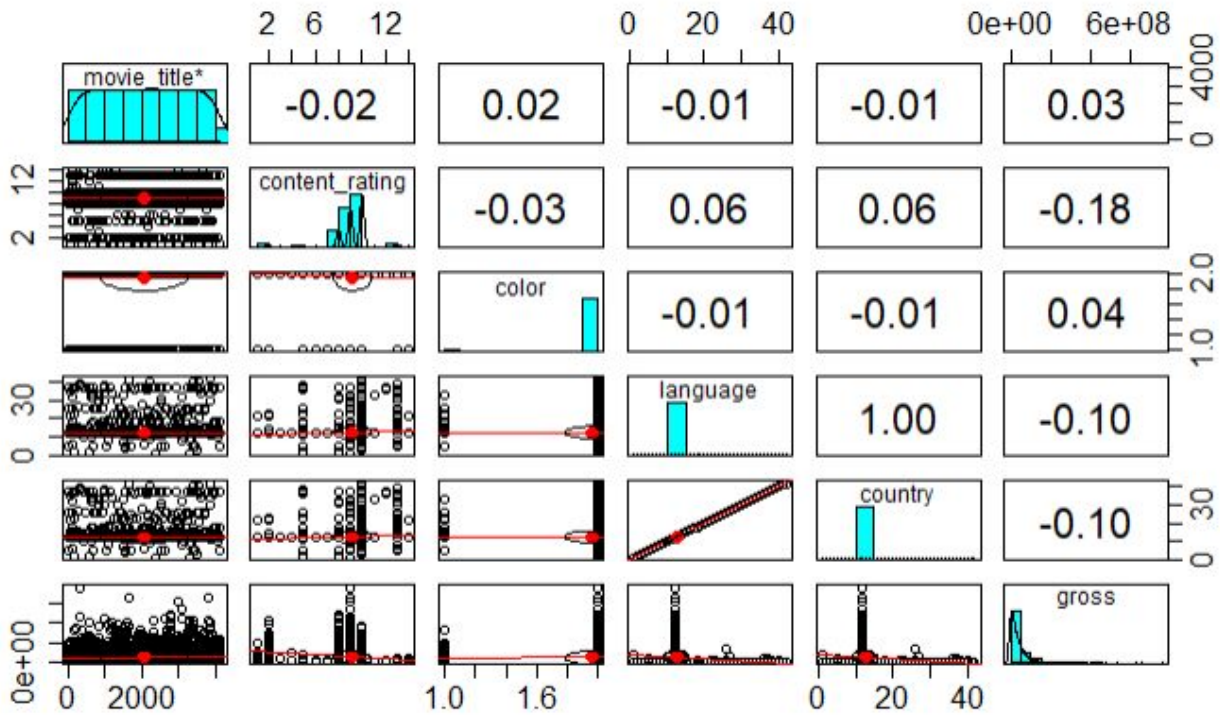
Correlation matrix

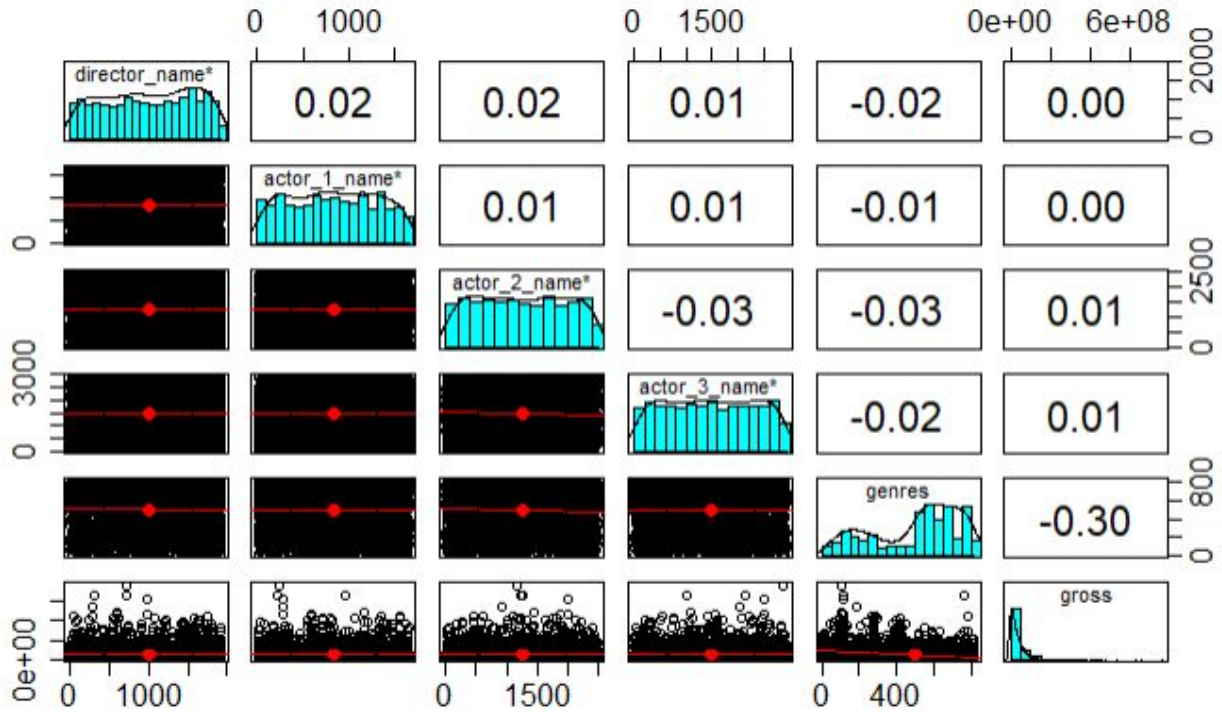
As we can see from correlation matrices as follows **(Figure 11**) , **num_user_for_reviews, num_voted_users, director_ave_gross, actor_1_ave_gross,actor_2_ave_gross, actor_3_ave_gross** has a moderate or strong positive correlation with gross. Among them, four of the newly added variable could be powerful powerful predictors. Other variables do not have an impact on the gross. It is surprising that imdb_score does not have the correlation to gross as expectation.

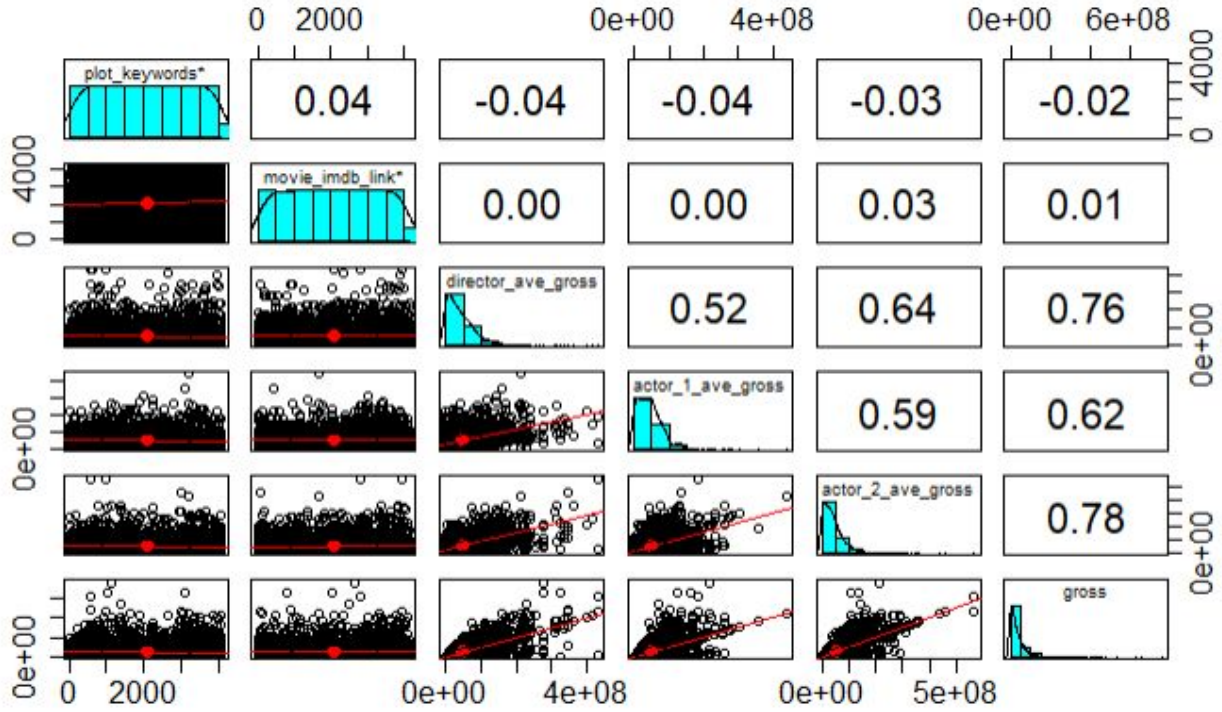

num_user_for_reviews
num_critic_for_reviews
imdb_score
num_voted_users
facenumber_in_poster
gross
0.58
0.30
0.79
-0.08
0.56
0.33
0.61
-0.03
0.48
0.45
-0.08
0.20
-0.03
0.63
0 400
0 1500000
0e+00 6e+08
0 800
0 5000
2 8
0 30
0e+00
0 3000
2 6
0 20 40


duration
aspect_ratio
budget
title_year
director_facebook_likes
gross
0.19
0.20
-0.13
0.17
0.24
0.14
0.32
0.05
0.09
0.13
0.06
0.38
-0.05
0.03
0.14
5 15
1920 2000
0e+00 6e+08
5
1920
50
0.0e+00
0 60000
0e+00
50 200
0.0e+00 2.0e+09
0 30000


actor_1_facebook_likes
actor_2_facebook_likes
actor_3_facebook_likes
cast_total_facebook_likes
movie_facebook_likes
gross
0.10
0.25
0.94
0.13
0.15
0.05
0.01
0.03
0.00
0.48
0.27
0.31
0.21
0.24
0.38


movie_title*
content_rating
color
language
country
gross
-0.02
0.02
-0.01
-0.01
0.03
-0.03
0.06
0.06
-0.18
-0.01
-0.01
0.04
1.00
-0.10
-0.10


director_name*
actor_1_name*
actor_2_name*
actor_3_name*
genres
gross
0.02
0.02
0.01
-0.02
0.00
0.01
0.01
-0.01
0.00
-0.03
-0.03
0.01
-0.02
0.01
-0.30
0 1000
0 1500
0e+00 6e+08
0 2000
0 2500
0 3000
0 800
0e+00
0 1000
0 1500
0 400


plot_keywords*
movie_imdb_link*
director_ave_gross
actor_1_ave_gross
actor_2_ave_gross
gross
0.04
-0.04
-0.04
-0.03
-0.02
0.00
0.00
0.03
0.01
0.52
0.64
0.76
0.59
0.62
0.78
0 2000
0e+00 4e+08
0e+00 6e+08
0 4000
0e+00
0 2000
0e+00 4e+08
0e+00 5e+08

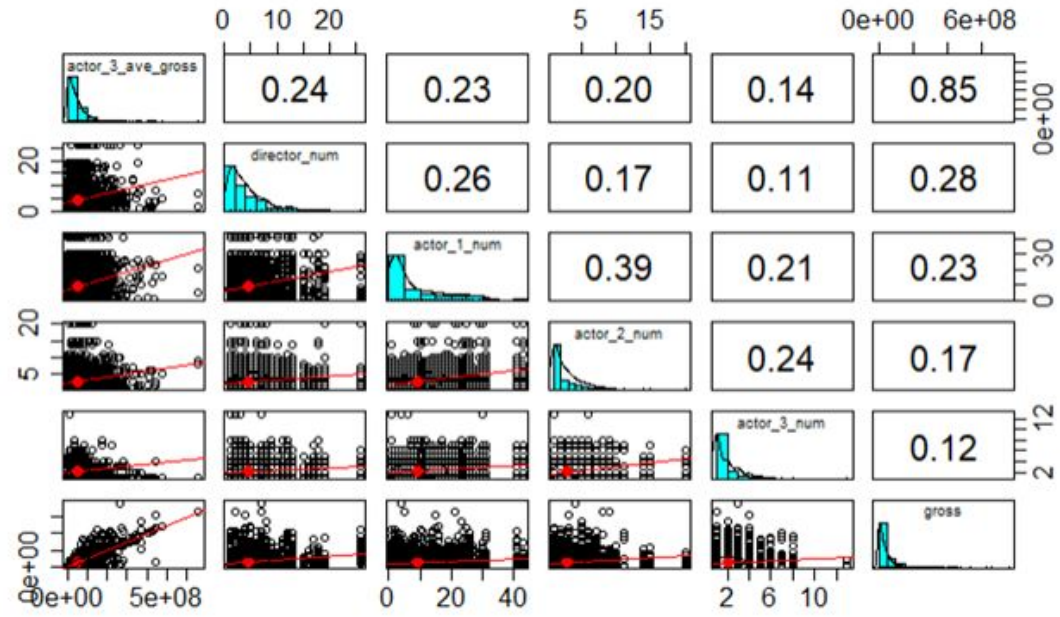


Figure 11. Correlation matrices.

The plots for log ***num_user_for_reviews,*** log ***num_voted_users,*** log ***director_ave_gross,*** log ***actor_1_ave_gross,*** log ***actor_2_ave_gross,*** log ***actor_3_ave_gross*** *and* log ***gross*** suggest that they have **linear** relationships.

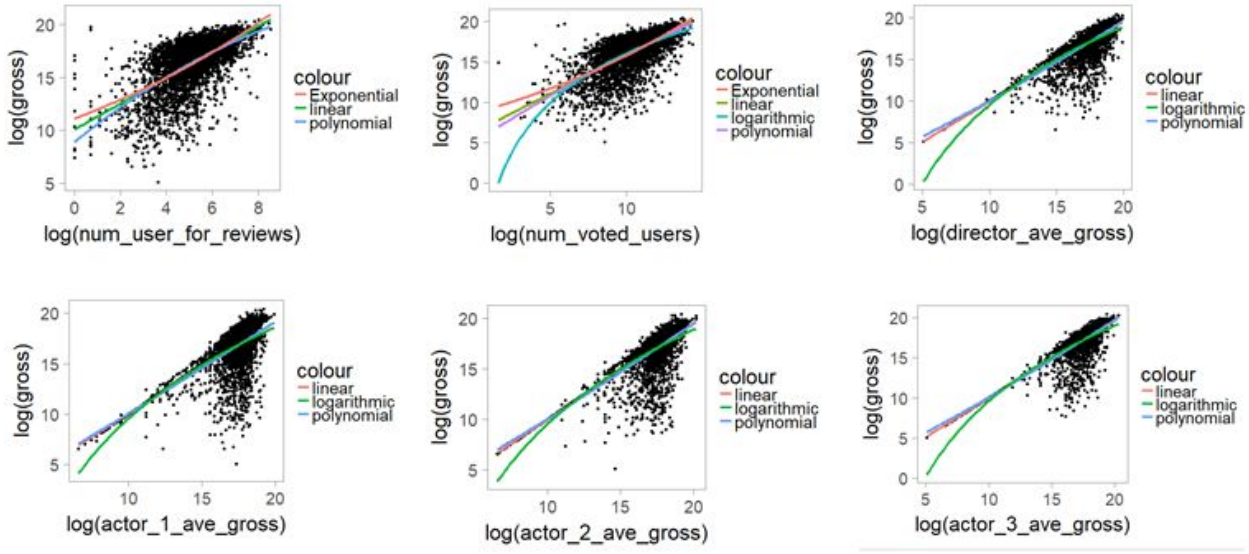


Figure 12. Correlation between 6 independent num_user_for_reviews, num_voted_users, director_ave_gross, actor_1_ave_gross, actor_2_ave_gross, actor_3_ave_gross and dependent variable gross.

VIF Test

Table3. VIF Test for Model-1

Variables <fctr>	VIF <dbl>
num_user_for_reviews	3.010056
num_critic_for_reviews	3.271651
imdb_score	1.489755
num_voted_users	4.000124
facenumber_in_poster	1.055173
duration	1.347187
budget	1.268979
title_year	1.406757
director_facebook_likes	1.551396
actor_1_facebook_likes	1.523236
actor_2_facebook_likes	1.184523
actor_3_facebook_likes	1.278070
movie_facebook_likes	2.226994
director_ave_gross	2.368348
actor_1_ave_gross	1.871656
actor_2_ave_gross	2.565049
actor_3_ave_gross	2.812658
director_num	1.604312
actor_1_num	1.724564
actor_2_num	1.259875
actor_3_num	1.100775

Variable inflation factor (VIF) test was performed to check the multiple collinearities between the dependent variables and 21-feature subsets were selected (**Table 3**). VIF test suggested that variable *cast_total_facebook_likes* should not be used because of the multicollinearity issue.

Outliers

Histogram

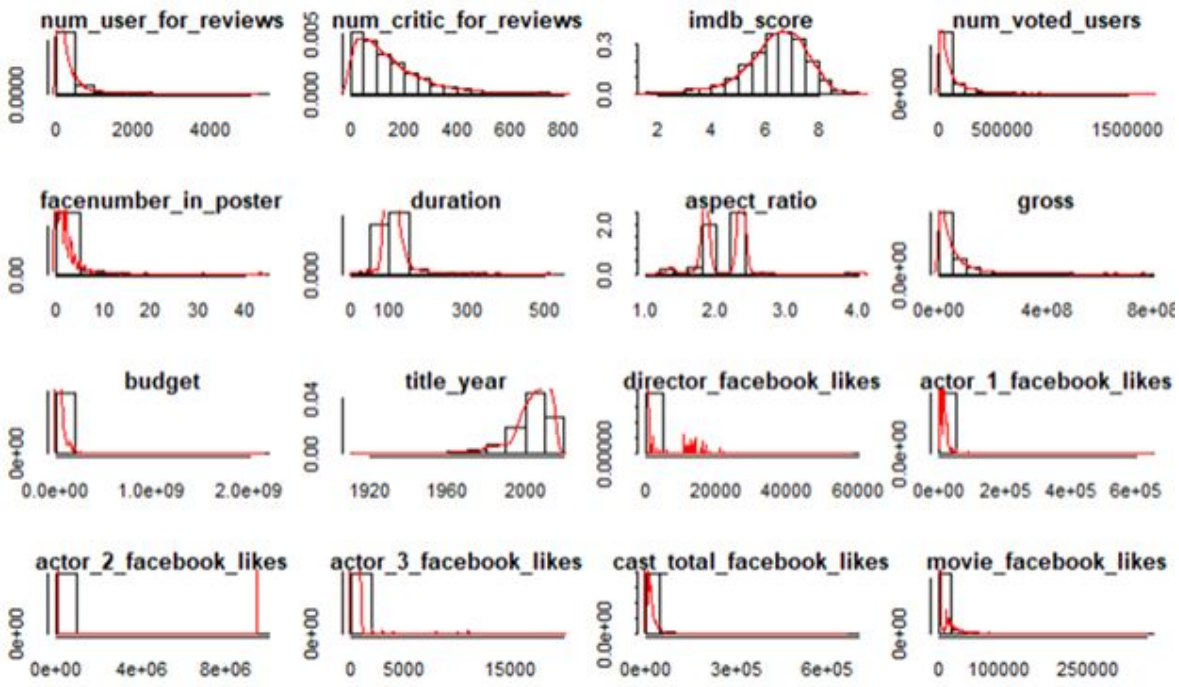


Figure 14. Hitogram of numeric variables.

Except for imdb_socre, all numeric variables are right-skewed. The plot of aspect_ratio suggests that it is not a continuous variable and we should treat it as a categorical variable.

Box plot

It looks like there are a lot of data points are beyond the 1.5 interquartile range in each numeric variable. But we do not consider them as outliers. Taking a random look at data reveals that these outliers are actual data points rather than experimental errors. Removing the data might not be a good approach risking the loss of information. Treatment of outliers might not produce accurate results as the imputation and score. We will keep the data as the way it is for now and do the outlier test in the multiple linear regression diagnoses.

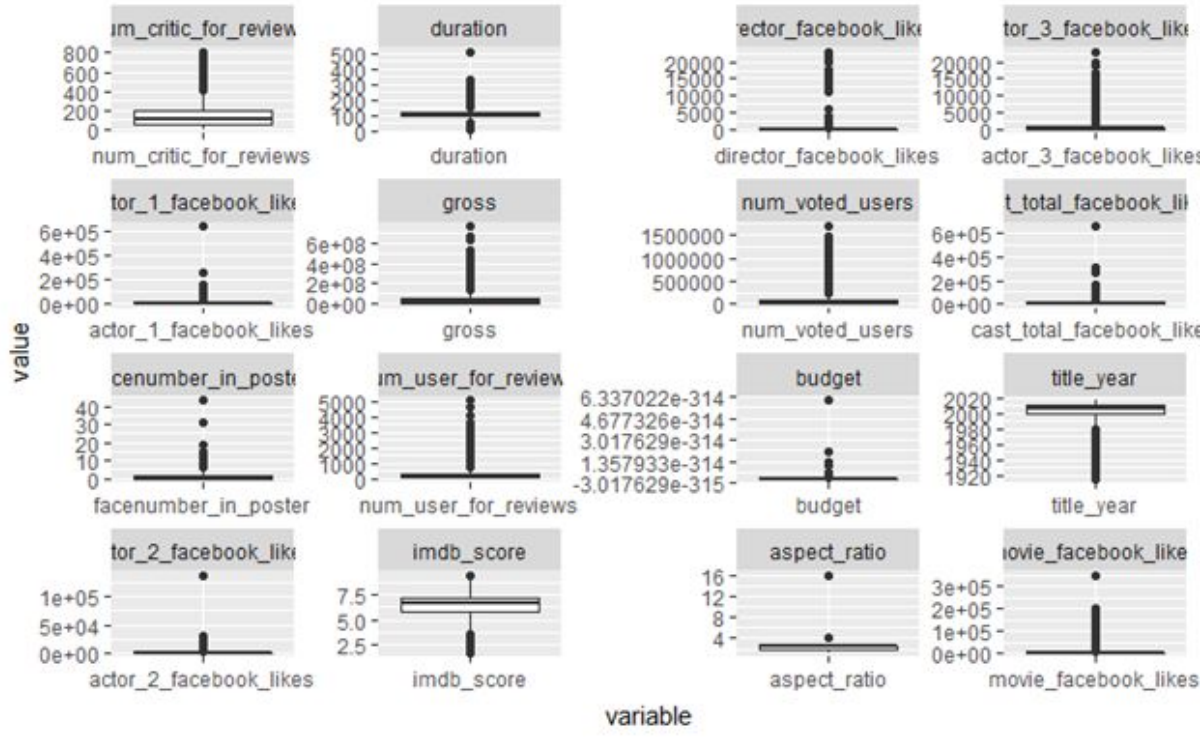


Figure 15. Box-plot of numeric variables.

Build Models

Model-1: Multiple Linear Regression I

Multiple linear regression model were built on the dataset with variables: "Index", "num_user_for_reviews", "num_critic_for_reviews", "imdb_score" ,"num_voted_users" , "facenumber_in_poster" "duration" , "aspect_ratio" , "gross", "budget","title_year" ,"director_facebook_likes", "actor_1_facebook_likes" , "actor_2_facebook_likes", "actor_3_facebook_likes", "cast_total_facebook_likes", "movie_facebook_likes", "movie_title", "content_rating" , "color", "language" , "country" , "director_name" , "actor_1_name", "actor_2_name", "actor_3_name", "genres", , "plot_keywords" , "movie_imdb_link" , "director_ave_gross", "actor_1_ave_gross" , , "actor_2_ave_gross" , "actor_3_ave_gross" , "director_num" , "actor_1_num" , "actor_2_num", "actor_3_num"

The outliers test indicate the records shown in the table below have Bonforonni p-value less than 0.5 (**Table 4)**.

Table 4. Outliers of Model-1

	rstudent	unadjusted p-value	Bonferonni p
1	14.522917	5.3036e-46	1.6245e-42
1231	-9.433231	8.5375e-21	2.6150e-17
30	7.550723	5.9611e-14	1.8259e-10
52	-5.979760	2.5425e-09	7.7878e-06
1395	5.530090	3.5209e-08	1.0785e-04
365	-5.354678	9.3239e-08	2.8559e-04
437	5.231786	1.8132e-07	5.5537e-04
4	5.134522	3.0386e-07	9.3074e-04
198	5.078471	4.0753e-07	1.2483e-03
562	-5.078471	4.0753e-07	1.2483e-03

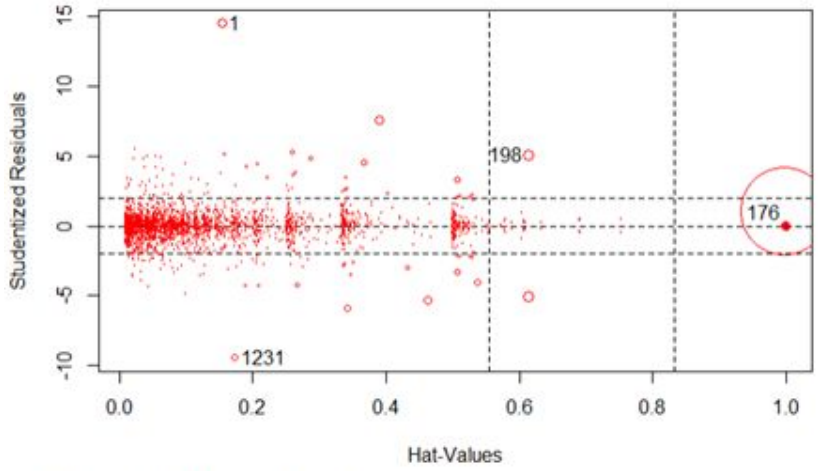


Figure 16. Influential points.

Removing outliers and influential point change magnitude or sign of coefficients of genres. The significant coefficients of the modified model are list **Table 5**. VIF test (**Table 6**) showed there was no collinearity in the final model.

Table 5. Coefficients of Model-1

	Estimate <dbl>	Std..Error <dbl>	t.value <dbl>	Pr...t.. <dbl>
num_voted_users	7.047237e+01	7.399735e+00	9.523635	3.695013e-21
aspect_ratio1.5	-9.401695e+07	3.778090e+07	-2.488478	1.289147e-02
aspect_ratio2	4.756703e+07	1.616566e+07	2.942474	3.284916e-03
budget	9.622645e-02	1.450376e-02	6.634587	3.947674e-11
director_facebook_likes	-9.082822e+02	1.991234e+02	-4.561404	5.317691e-06
actor_1_facebook_likes	-1.111174e+02	4.782201e+01	-2.323561	2.022603e-02
actor_3_facebook_likes	-1.371376e+03	3.173784e+02	-4.320951	1.612140e-05
content_ratingR	-1.252972e+07	5.640698e+06	-2.221306	2.641634e-02
colorColor	1.027690e+07	3.031957e+06	3.389527	7.106124e-04
genresAction|Adventure|Comedy|Family|Fantasy	7.569083e+07	2.879614e+07	2.628506	8.626719e-03
genresAction|Adventure|Comedy|Family|Fantasy|Sci_Fi	-5.770279e+07	1.954507e+07	-2.952294	3.182563e-03
genresAction|Adventure|Crime|Drama|Mystery|Thriller	6.197611e+07	2.876825e+07	2.154323	3.130677e-02
genresAction|Adventure|Drama|History|War	-4.011643e+07	1.627860e+07	-2.464365	1.379008e-02
genresAction|Adventure|Drama|Sci_Fi|Thriller	-8.963541e+07	3.364262e+07	-2.664341	7.761741e-03
genresAction|Adventure|Family|Sci_Fi	-4.833490e+07	2.223793e+07	-2.173535	2.983039e-02
genresAction|Adventure|Fantasy|Romance	9.761742e+07	1.951739e+07	5.001562	6.067031e-07
genresAction|Adventure|Fantasy|Sci_Fi	4.723316e+07	1.508717e+07	3.130684	1.763434e-03
genresAction|Adventure|Horror|Sci_Fi	-4.007143e+07	1.789570e+07	-2.239165	2.522971e-02
genresAction|Adventure|Romance|Sci_Fi|Thriller	-8.916736e+07	2.220761e+07	-4.015172	6.108999e-05
genresAction|Animation|Comedy|Family|Sci_Fi	7.336906e+07	2.246493e+07	3.265938	1.105239e-03
genresAction|Animation|Sci_Fi	-1.084037e+08	3.428372e+07	-3.161958	1.585239e-03
genresAction|Biography|Drama|History|Romance|War	-6.003502e+07	2.878471e+07	-2.085657	3.710736e-02
genresAction|Comedy|Romance|Thriller	-4.724267e+07	2.237128e+07	-2.111755	3.480298e-02
genresAction|Drama|Mystery|Sci_Fi	-5.917653e+07	2.887617e+07	-2.049320	4.053131e-02
genresAction|Western	-7.392241e+07	2.867803e+07	-2.577667	1.000177e-02
genresAdventure|Animation|Comedy|Family|Sport	5.720151e+07	2.283301e+07	2.505211	1.229875e-02
genresAdventure|Animation|Drama|Family|Fantasy	-1.032687e+08	2.894023e+07	-3.568343	3.657653e-04
genresAdventure|Animation|Drama|Family|Musical	1.004262e+08	3.005945e+07	3.340918	8.468584e-04
genresAdventure|Comedy|Family|Mystery|Sci_Fi	9.729601e+07	2.879525e+07	3.378891	7.385537e-04
genresAdventure|Comedy|Fantasy	-4.734159e+07	2.215450e+07	-2.136883	3.270087e-02
genresAdventure|Comedy|Fantasy|Sci_Fi	-8.938838e+07	2.895855e+07	-3.086770	2.044846e-03
genresAdventure|Drama	-3.205044e+07	1.632394e+07	-1.963401	4.970641e-02
genresAdventure|Drama|Fantasy|Romance	4.439784e+07	1.980703e+07	2.241520	2.507678e-02
genresAdventure|Fantasy|Mystery	-6.249960e+07	2.869722e+07	-2.177897	2.950358e-02
genresAnimation|Comedy|Family|Fantasy|Music	4.827757e+07	2.230437e+07	2.164489	3.051789e-02
genresAnimation|Drama|Family|Fantasy|Musical|Romance	-6.195227e+07	2.894124e+07	-2.140622	3.239758e-02
genresBiography|Comedy|Crime|Drama	-4.688229e+07	1.944522e+07	-2.410993	1.597832e-02
genresBiography|Crime|Drama|History|Western	-6.257742e+07	2.884068e+07	-2.169762	3.011546e-02
genresBiography|Crime|Drama|Romance|Thriller	-6.464115e+07	2.878163e+07	-2.245917	2.479330e-02
genresBiography|Drama|History|Romance	-3.671502e+07	1.589788e+07	-2.309428	2.099811e-02
genresComedy|Crime|Drama|Mystery	-9.196786e+07	3.227069e+07	-2.849889	4.408010e-03
genresComedy|Horror|Sci_Fi	-6.388903e+07	2.879038e+07	-2.219110	2.656555e-02
genresDrama | Fantasy | Thriller	-1.592910e+08	3.168131e+07	-5.027915	5.296954e-07
genresDrama|Fantasy|Thriller	-5.373648e+07	1.946535e+07	-2.760623	5.809758e-03
genresDrama|Musical	-4.300540e+07	2.037422e+07	-2.110775	3.488721e-02
genresFamily|Sci_Fi	2.231789e+08	2.901797e+07	7.691057	2.057892e-14
genresFantasy | Horror | Mystery	-9.581513e+07	3.175918e+07	-3.016928	2.578344e-03
director_ave_gross	2.425076e-01	1.619826e-02	14.971215	1.156310e-48
actor_1_ave_gross	1.498446e-01	1.657780e-02	9.038870	3.021005e-19
actor_2_ave_gross	3.194067e-01	1.548762e-02	20.623352	1.267887e-87
actor_3_ave_gross	4.075478e-01	1.537190e-02	26.512519	3.025650e-137
actor_1_num	-1.882114e+05	6.830025e+04	-2.755647	5.898603e-03

Table 6. VIF Test for Modified Model-1 (Final Model)

Variables <fctr>	VIF <dbl>
num_voted_users	2.369049
budget	3.557834
director_facebook_likes	1.344705
actor_1_facebook_likes	1.725678
actor_3_facebook_likes	1.540898
color	*NA*
genres	*NA*
director_ave_gross	3.225087
actor_1_ave_gross	2.372722
actor_2_ave_gross	3.455505
actor_3_ave_gross	3.738799
actor_1_num	2.018039

Prediction

The model could only predict 17 gross values out of 849 records (**Table 6**) while the other predicted value is missing (NA). The predicted box office only showed consistency in case 924 and 16. Quite a big part of the prediction is negative. Thus, a multiple linear model might not be suitable for the data.

Table 6. Predicted vs. Actual Gross (Model-1)

	Index <dbl>	gross <dbl>	pred <dbl>
179	179	83024900	139376071
2605	2605	138795342	92594019
1158	1158	57750000	9976874
1065	1065	92001027	-18190851
694	694	19548064	104536308
4955	4955	110536	-77859658
3156	3156	11703287	-26328934
924	924	296623634	301865703
2597	2597	7774730	-32432633
235	235	310675583	407445950
2133	2133	42660000	-9079887
4788	4788	1229197	-2636808
379	379	61355436	-22623777
2805	2805	2474000	-30706632
1356	1356	10300000	-1046228
16	16	291021565	275958106
1000	1000	26616999	-57427000

Model-2: Multiple Linear Regression II

The same variables to model-1 have been used for the multiple linear regression models except that the categorical variable “genres”, “color”, “content_rating”, “aspect_ratio”, “language” and “country” have been dummy encoded. Model-2 has much less significant coefficients comparing to model-1 (less than half) as shown in **Table 7**.

Table 7. Coefficients for Model-2

	Estimate <dbl>	Std..Error <dbl>	t.value <dbl>	Pr...t.. <dbl>
(Intercept)	2.742317e+08	1.382791e+08	1.983176	4.743100e-02
num_voted_users	7.220259e+01	6.729021e+00	10.730029	2.007113e-26
budget	4.380149e-02	8.841787e-03	4.953918	7.640960e-07
title_year	-1.337406e+05	6.683453e+04	-2.001071	4.546697e-02
director_facebook_likes	-8.061407e+02	1.882712e+02	-4.281806	1.906911e-05
actor_3_facebook_likes	-1.205453e+03	4.318249e+02	-2.791531	5.276271e-03
director_ave_gross	2.638018e-01	1.427532e-02	18.479564	1.286648e-72
actor_1_ave_gross	1.593852e-01	1.509264e-02	10.560457	1.166655e-25
actor_2_ave_gross	3.188490e-01	1.391196e-02	22.919064	5.816603e-108
actor_3_ave_gross	4.386336e-01	1.373677e-02	31.931347	4.399935e-195
actor_1_num	-1.999891e+05	6.276820e+04	-3.186154	1.455301e-03
actor_2_num	-4.271554e+05	2.043329e+05	-2.090488	3.665075e-02
genre.Fantasy	-3.511051e+06	1.567339e+06	-2.240135	2.514883e-02
genre.Family	8.278381e+06	2.457703e+06	3.368341	7.649363e-04
genre.Drama	-2.657662e+06	1.241260e+06	-2.141100	3.233943e-02
genre.History	-5.715232e+06	2.762160e+06	-2.069117	3.861320e-02
color.Color	6.575856e+06	2.693947e+06	2.440974	1.470020e-02
content_ratingApproved	-2.043656e+07	9.450724e+06	-2.162434	3.065675e-02
content_ratingG	-1.202543e+07	6.015377e+06	-1.999114	4.567831e-02
content_ratingR	-1.331142e+07	4.919119e+06	-2.706057	6.843944e-03
aspect1.5	-7.576946e+07	3.019944e+07	-2.508969	1.215622e-02

Prediction

The predicted *vs.* actual box office is shown in **Figure 17**. Several link functions were fit to see whether changing the link function could improve the fitting. It seems none of them can perfectly fit the data.

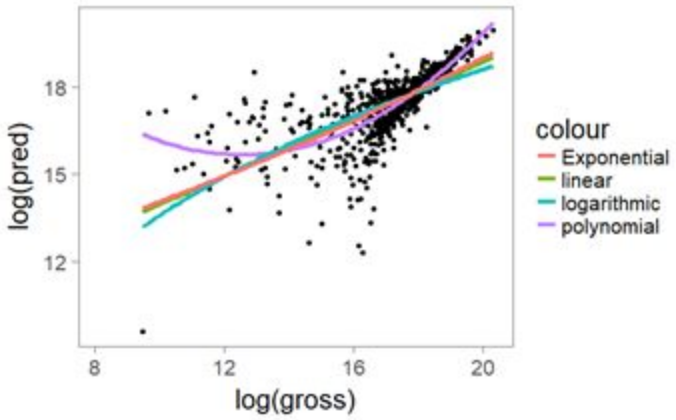


Figure 17. Predicted vs. actual gross in model-2.

In the rest of this report, we will put the box office into 10 buckets, by sections on a scale (0~1m, 1m~3m, 3m~6.25m,6.25m~12.5m, 12.5m~25m, 25m~50m, 50m~100m, 100m~300m, 500m+), and fit the data with ordinal regression. The number of each class in the evaluation set is tallied as in **Table 8**.

Table 8. Classes of Gross

Class <fctr>	Range <fctr>	Frequancy <fctr>
1	0~1m	114
2	1m~3m	56
3	3m~6.25m	63
4	6.25m~12.5m	76
5	12.5m~25m	128
6	25m~50m	141
7	50m~100m	143
8	100m~300m	110
9	300m~500m	15
10	500m+	3

The Generalized Linear Regression, Ordinal Logistic Regression, and Classification Trees models were built for this movie dataset with an ordinal dependent variable.

Model-3: Ordinal Logistic Regression (OLR)

A Ordinal Logistic Regression model was built with vglm function from VGAM package. We first tried the parallel regression assumption and got a negative infinitive Log Likelihood, suggesting the Log Likelihood reached the boundaries of the parameter space. So we built a partial proportional odds model. By using a partial proportional odds model, we build a mix of ordinal and multinomial logistic regression.

The categorical variables genres, color, content_rating, aspect_ratio were transformed to dummy variables. In order to build the model without getting an error of NA or Inf value during the computation, some categorical variables with too many levels including *language, country, movie_title, director_name, actor_1_name, actor_2_name, actor_3_name, plot_keywords, movie_imdb_link* were dropped. In addition, some of the nested levels with very low frequency including *genre.News, genre.Short, content_ratingTV_MA, content_ratingTV_PG, content_ratingGP, content_ratingM,content_ratingPassed, aspect1.44, aspect1.5, aspect1.75, aspect1.77, aspect2.24, aspect2.4, aspect2.55* and *aspect2.76* were dropped. One continuous variable *imdb_score* was also dropped.

Prediction

The OLR model were used to predict the box office for the evaluation set. The confusion matrix and the statistics by class are shown as follows. The prediction only showed reasonable accuracy for class-1 and class-9. The overall accuracy is only 37%.

```
          Reference
Prediction   1   2   3   4   5   6   7   8   9  10
        1  100   0  40  32  38  35  27   8   0   0
        2    2   0   0   3   0   1   1   0   0   0
        3    0   0   1   2   1   0   0   0   0   0
        4    0   0   2   2   2   0   0   0   0   0
        5    1   0   3   4   5  17   8   3   0   0
        6   10   0   8  18  22  58  60  31   0   0
        7    0   0   0   0   3  15  29  24   6   0
        8    0   0   0   1   4   1   9  47  32   0
        9    1   3   2   1   1   1   7  30  72  15
        10   0   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.3698
                 95% CI : (0.3373, 0.4033)
    No Information Rate : 0.1684
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2671
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7
Sensitivity            0.8772 0.000000 0.017857 0.031746 0.065789  0.45312  0.20567
Specificity            0.7551 0.991726 0.996217 0.994911 0.953428  0.79334  0.93220
Pos Pred Value         0.3571 0.000000 0.250000 0.333333 0.121951  0.28019  0.37662
Neg Pred Value         0.9754 0.996437 0.934911 0.927639 0.912129  0.89097  0.85492
Prevalence             0.1343 0.003534 0.065960 0.074205 0.089517  0.15077  0.16608
Detection Rate         0.1178 0.000000 0.001178 0.002356 0.005889  0.06832  0.03416
Detection Prevalence   0.3298 0.008245 0.004711 0.007067 0.048292  0.24382  0.09069
Balanced Accuracy      0.8161 0.495863 0.507037 0.513328 0.509609  0.62323  0.56894
                     Class: 8 Class: 9 Class: 10
Sensitivity           0.32867  0.65455   0.00000
Specificity           0.93343  0.91746   1.00000
Pos Pred Value        0.50000  0.54135       NaN
Neg Pred Value        0.87285  0.94693   0.98233
Prevalence            0.16843  0.12956   0.01767
Detection Rate        0.05536  0.08481   0.00000
Detection Prevalence  0.11072  0.15665   0.00000
Balanced Accuracy     0.63105  0.78600   0.50000
```

Model-4: Random Forest (RF)

A random forest model was built using the same dataset as what model-2 used. The importance of the 30 significant predictors is ranked (**Figure 18**). The most weighted predictor is the average box office of the movies actor_3 previously involved in. The second, third and fourth important predictor are the average box office of the movies the director and other two actors previously involved in.

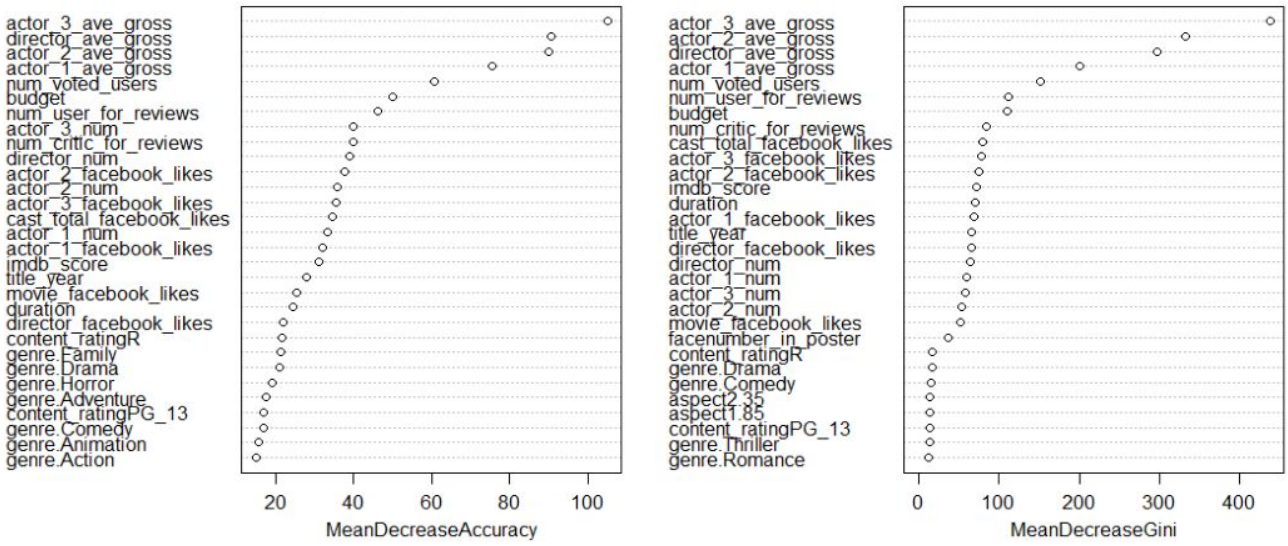


Figure 18. The importance of predictors in RF model.

Prediction

The RF model was used to predict the box office. The confusion matrix is shown as follows. The accuracy is 77%, which is much higher than model-1, model-2, and model-3.

```
          Reference
Prediction   1   2   3   4   5   6   7   8   9  10
        1  103   0   4   1   1   0   0   0   0   0
        2    0   1   0   0   0   0   0   0   0   0
        3    1   0  38   1   0   0   0   0   0   0
        4    0   0   4  51   1   0   0   0   0   0
        5    0   0   2   3  36   2   0   0   0   0
        6    3   0   7   4  25  94   5   0   0   0
        7    7   0   1   3  12  27 122  22   0   0
        8    0   0   0   0   1   5  14 109  15   0
        9    0   1   0   0   0   0   0  12  95   8
        10   0   1   0   0   0   0   0   0   0   7

Overall Statistics

               Accuracy : 0.7727
                 95% CI : (0.743, 0.8005)
    No Information Rate : 0.1684
    P-Value [Acc > NIR] : < 2.2e-16
```

The statistic for each class from class-1 to class-10 is summarized in the table below. Except for class-2, class-4, and class-10, the sensitivity is greater than 67%. The specificity of each class is high, 89% ~ 99%.

Statistics by Class:

	Class: 1	Class: 2	Class: 3	Class: 4	Class: 5	Class: 6	Class: 7
Sensitivity	0.9035	0.333333	0.67857	0.80952	0.47368	0.7344	0.8652
Specificity	0.9918	1.000000	0.99748	0.99364	0.99094	0.9390	0.8983
Pos Pred Value	0.9450	1.000000	0.95000	0.91071	0.83721	0.6812	0.6289
Neg Pred Value	0.9851	0.997642	0.97775	0.98487	0.95037	0.9522	0.9710
Prevalence	0.1343	0.003534	0.06596	0.07420	0.08952	0.1508	0.1661
Detection Rate	0.1213	0.001178	0.04476	0.06007	0.04240	0.1107	0.1437
Detection Prevalence	0.1284	0.001178	0.04711	0.06596	0.05065	0.1625	0.2285
Balanced Accuracy	0.9477	0.666667	0.83802	0.90158	0.73231	0.8367	0.8818

	Class: 8	Class: 9	Class: 10
Sensitivity	0.7622	0.8636	0.466667
Specificity	0.9504	0.9716	0.998801
Pos Pred Value	0.7569	0.8190	0.875000
Neg Pred Value	0.9518	0.9795	0.990488
Prevalence	0.1684	0.1296	0.017668
Detection Rate	0.1284	0.1119	0.008245
Detection Prevalence	0.1696	0.1366	0.009423
Balanced Accuracy	0.8563	0.9176	0.732734

Model-5: Generalized Linear Models with L1 Penalty (GLM)

The glmpathcr function from glmpathcr package was used to build the Generalized Linear Models model because glmpathcr is a frame designed for the high dimensional dataset. This method uses a path-following algorithm for L1 regularized generalized linear models. The L1 regularization selects variables according to the amount of penalization on the L1 norm of the coefficients, in a manner less greedy than forwarding selection/backward deletion.

The number of non-zero coefficients under AIC criteria is shown below:

Intercept	num_user_for_reviews	num_critic_for_reviews
5.326597e+01	2.197342e-04	4.331548e-03
imdb_score	num_voted_users	facenumber_in_poster
-1.276490e-01	3.234294e-06	1.088422e-02
duration	budget	title_year
7.540256e-03	5.526017e-11	-3.228145e-02
director_facebook_likes	actor_1_facebook_likes	actor_2_facebook_likes
-4.931502e-05	6.774942e-06	2.898404e-07
actor_3_facebook_likes	cast_total_facebook_likes	movie_facebook_likes
-4.046599e-05	-1.322444e-05	-8.644571e-06
director_ave_gross	actor_1_ave_gross	actor_2_ave_gross
1.570583e-08	8.360939e-09	1.571561e-08
actor_3_ave_gross	director_num	actor_2_num
2.452833e-08	9.009791e-03	-4.644004e-02
actor_3_num	genre.Action	genre.Adventure
-9.806932e-02	2.890595e-01	-5.995831e-02
genre.Fantasy	genre.Sci_Fi	genre.Thriller
-2.362776e-01	-2.433909e-01	1.575178e-01
genre.Romance	genre.Animation	genre.Comedy
1.734597e-01	1.286542e-01	5.892585e-02
genre.Family	genre.Musical	genre.Mystery
6.751232e-01	-2.952141e-01	1.975672e-01
genre.Western	genre.Drama	genre.History
-1.280554e-01	-2.565208e-01	-2.585655e-01
genre.Sport	genre.Crime	genre.Horror
3.403801e-01	-9.160139e-02	1.874011e-01
genre.War	genre.Biography	genre.Music
-7.105685e-03	-1.998613e-02	4.617875e-01
genre.Documentary	genre.News	genre.Short
4.538491e-01	8.167509e-01	-2.648930e+00
color.Color	content_ratingApproved	content_ratingG
6.697075e-01	2.952433e-01	9.220431e-01
content_ratingM	content_ratingPassed	content_ratingPG
1.396728e-01	-5.794635e-01	1.040213e+00
content_ratingPG_13	content_ratingR	content_ratingTV_MA
1.221243e+00	6.461700e-01	-2.322000e+00
content_ratingTV_PG	content_ratingUnrated	content_ratingX
2.197310e+00	-2.575508e-01	-1.826001e-01
aspect1.33	aspect1.37	aspect1.44
-2.445470e-01	-3.386284e-01	1.081537e+00
aspect1.5	aspect1.66	aspect1.75
-2.445761e+00	1.271932e-01	-4.018608e+00
aspect1.77	aspect1.78	aspect1.85
-3.152183e-01	2.017220e-02	8.448194e-02
aspect2	aspect2.2	aspect2.39
-7.898016e-02	-1.018742e+00	-2.268378e-01
aspect2.4	aspect2.55	aspect2.76
1.085280e+00	-1.700711e+00	-3.018783e-01
cp1	cp2	cp3
8.322670e+00	7.779721e+00	7.345358e+00
cp4	cp5	cp6
7.169964e+00	6.475499e+00	4.882875e+00
cp7	cp8	cp9
2.517785e+00	-6.236148e+00	-1.655555e+01

The number of non-zero coefficients under BIC criteria is less:

Intercept	num_user_for_reviews	num_critic_for_reviews
2.442227e+01	1.007765e-04	2.537296e-03
imdb_score	num_voted_users	duration
-1.168802e-05	5.661668e-07	2.420727e-03
title_year	director_facebook_likes	actor_3_facebook_likes
-1.495386e-02	-1.233291e-05	-7.005783e-05
cast_total_facebook_likes	director_ave_gross	actor_1_ave_gross
-2.920190e-06	1.166697e-08	7.730280e-09
actor_2_ave_gross	actor_3_ave_gross	director_num
1.180623e-08	1.427293e-08	1.218900e-02
actor_2_num	actor_3_num	genre.Action
-3.968105e-03	-6.610188e-03	1.044200e-01
genre.Thriller	genre.Family	genre.Mystery
1.144224e-02	4.070072e-01	3.565490e-02
genre.Drama	genre.Music	color.Color
-2.850861e-01	4.585954e-02	2.043848e-01
content_ratingPG	content_ratingPG_13	content_ratingUnrated
2.069714e-01	3.448795e-01	-5.678454e-01
cp1	cp2	cp3
-4.547814e+00	3.570141e+00	3.148231e+00
cp4	cp5	cp6
2.842220e+00	2.819717e+00	2.391898e+00
cp7	cp8	cp9
1.206833e+00	-4.381215e-01	-6.252595e+00

Prediction

Using the GLM to do a prediction for evaluation set and the confusion matrix is shown as follow. AIC is used as criteria for the feature selection. The overall accuracy is 49%. No movie was assigned to class-2 to class-4. In other classes, the sensitivity is between 49% to 83%.

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3  4  5  6  7  8  9 10
        1  95 39 35 25 10  1  0  0  0  0
        2   0  0  0  0  0  0  0  0  0  0
        3   0  0  0  0  0  0  0  0  0  0
        4   0  0  0  0  0  0  0  0  0  0
        5  15 14 23 33 63 23  3  0  0  0
        6   3  2  5 16 50 96 47  4  0  0
        7   1  1  0  1  5 15 78 28  0  0
        8   0  0  0  1  0  6 15 77  4  0
        9   0  0  0  0  0  0  0  1 10  1
        10  0  0  0  0  0  0  0  0  1  2

Overall Statistics

               Accuracy : 0.4959
                 95% CI : (0.4617, 0.5301)
    No Information Rate : 0.1684
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4079
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Sensitivity            0.8333  0.00000   0.0000  0.00000   0.4922   0.6809  0.54545  0.70000
Specificity            0.8503  1.00000   1.0000  1.00000   0.8460   0.8206  0.92776  0.96482
Pos Pred Value         0.4634      NaN      NaN      NaN   0.3621   0.4305  0.60465  0.74757
Neg Pred Value         0.9705  0.93404   0.9258  0.91048   0.9037   0.9281  0.90972  0.95576
Prevalence             0.1343  0.06596   0.0742  0.08952   0.1508   0.1661  0.16843  0.12956
Detection Rate         0.1119  0.00000   0.0000  0.00000   0.0742   0.1131  0.09187  0.09069
Detection Prevalence   0.2415  0.00000   0.0000  0.00000   0.2049   0.2627  0.15194  0.12132
Balanced Accuracy      0.8418  0.50000   0.5000  0.50000   0.6691   0.7507  0.73661  0.83241
                     Class: 9 Class: 10
Sensitivity           0.66667  0.666667
Specificity           0.99760  0.998818
Pos Pred Value        0.83333  0.666667
Neg Pred Value        0.99403  0.998818
Prevalence            0.01767  0.003534
Detection Rate        0.01178  0.002356
Detection Prevalence  0.01413  0.003534
Balanced Accuracy     0.83213  0.832742
```

When BIC was employed, the result below showed that no movie was assigned to class-9 and class-10. For the rest classes, it seemed the model estimated one class higher for all the classes except class-1. And the accuracy is worse than AIC which is only 22%.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7   8   9  10
        1  104   0  10   2   1   2   0   0   0   0
        2    1   0  34   1   0   0   0   0   0   0
        3    0   0   3  45   4   1   0   0   0   0
        4    0   0   2   3  45   9   0   0   0   0
        5    5   0   4   9  12  68  31   4   0   0
        6    4   0   3   2   9  38  90  35   3   0
        7    0   0   0   1   5  10  20  96  52   0
        8    0   3   0   0   0   0   0   8  55  15
        9    0   0   0   0   0   0   0   0   0   0
        10   0   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.2214
                 95% CI : (0.1939, 0.2509)
    No Information Rate : 0.1684
    P-Value [Acc > NIR] : 4.039e-05

                  Kappa : 0.1082
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7
Sensitivity            0.9123 0.000000 0.053571 0.047619  0.15789  0.29688  0.14184
Specificity            0.9796 0.957447 0.936948 0.928753  0.84347  0.79750  0.76836
Pos Pred Value         0.8739 0.000000 0.056604 0.050847  0.09023  0.20652  0.10870
Neg Pred Value         0.9863 0.996310 0.933417 0.924051  0.91061  0.86466  0.81805
Prevalence             0.1343 0.003534 0.065960 0.074205  0.08952  0.15077  0.16608
Detection Rate         0.1225 0.000000 0.003534 0.003534  0.01413  0.04476  0.02356
Detection Prevalence   0.1402 0.042403 0.062426 0.069494  0.15665  0.21673  0.21673
Balanced Accuracy      0.9459 0.478723 0.495260 0.488186  0.50068  0.54719  0.45510
                     Class: 8 Class: 9 Class: 10
Sensitivity          0.055944   0.0000   0.00000
Specificity          0.896601   1.0000   1.00000
Pos Pred Value       0.098765      NaN       NaN
Neg Pred Value       0.824219   0.8704   0.98233
Prevalence           0.168433   0.1296   0.01767
Detection Rate       0.009423   0.0000   0.00000
Detection Prevalence 0.095406   0.0000   0.00000
Balanced Accuracy    0.476272   0.5000   0.50000
```

Model-6: Classification Trees

Rpart function was used to fit a classification tree for the train set(Figure 19). Similar to the Random Forest Tree model, the average box office made by the previous movies that the director and actors involved are the main features that the model used to build the regression tree and the actor_3_ave_gross sits on the tree top.

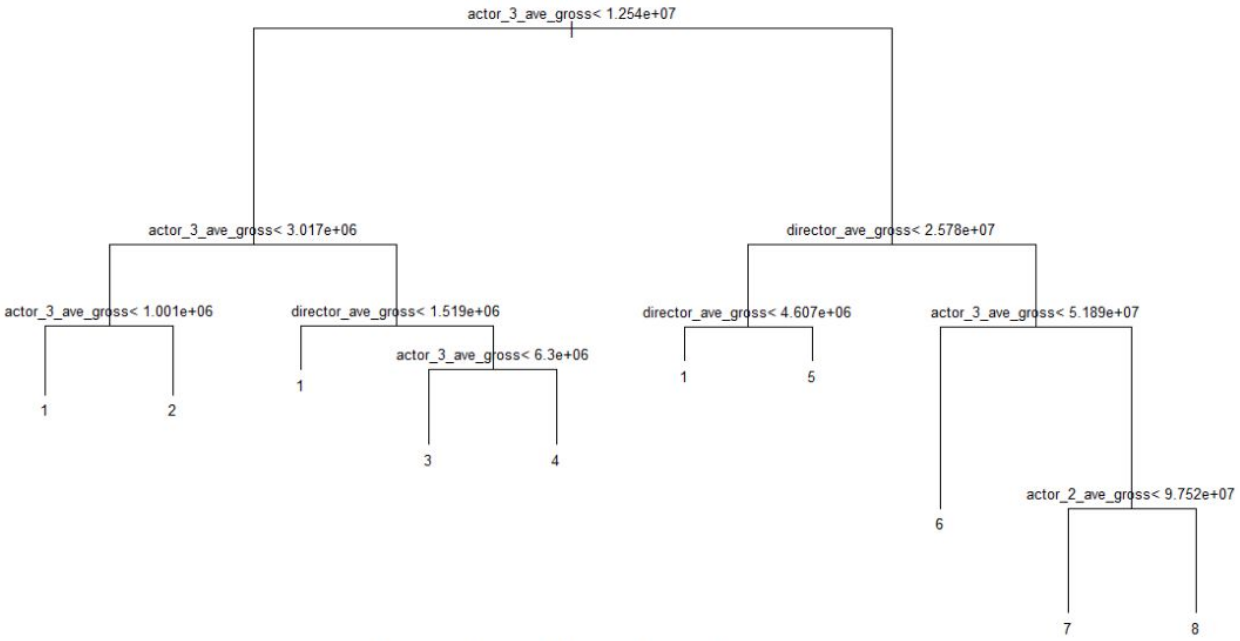


Figure 19. The Classification Trees model. cp= 0.01.

Prediction

The Classification Trees model was used to do a prediction for the evaluation set and the confusion matrix is as follows. The overall accuracy is 63% and there is no movie assigned to class-9 and class-10. The sensitivity for each class is between 50% and 91%.

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7   8   9  10
        1  104  10   2   1   2   0   0   0   0   0
        2    1  34   1   0   0   0   0   0   0   0
        3    0   3  45   4   1   0   0   0   0   0
        4    0   2   3  45   9   0   0   0   0   0
        5    5   4   9  12  68  31   4   0   0   0
        6    4   3   2   9  38  90  35   3   0   0
        7    0   0   1   5  10  20  96  52   0   0
        8    0   0   0   0   0   0   8  55  15   3
        9    0   0   0   0   0   0   0   0   0   0
        10   0   0   0   0   0   0   0   0   0   0

Overall Statistics

               Accuracy : 0.6325
                 95% CI : (0.5991, 0.665)
    No Information Rate : 0.1684
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5722
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6 Class: 7 Class: 8
Sensitivity            0.9123  0.60714  0.71429  0.59211  0.53125   0.6383   0.6713  0.50000
Specificity            0.9796  0.99748  0.98982  0.98189  0.90985   0.8672   0.8754  0.96482
Pos Pred Value         0.8739  0.94444  0.84906  0.76271  0.51128   0.4891   0.5217  0.67901
Neg Pred Value         0.9863  0.97294  0.97739  0.96076  0.91620   0.9233   0.9293  0.92839
Prevalence             0.1343  0.06596  0.07420  0.08952  0.15077   0.1661   0.1684  0.12956
Detection Rate         0.1225  0.04005  0.05300  0.05300  0.08009   0.1060   0.1131  0.06478
Detection Prevalence   0.1402  0.04240  0.06243  0.06949  0.15665   0.2167   0.2167  0.09541
Balanced Accuracy      0.9459  0.80231  0.85205  0.78700  0.72055   0.7528   0.7733  0.73241
                     Class: 9 Class: 10
Sensitivity           0.00000  0.000000
Specificity           1.00000  1.000000
Pos Pred Value            NaN       NaN
Neg Pred Value        0.98233  0.996466
Prevalence            0.01767  0.003534
Detection Rate        0.00000  0.000000
Detection Prevalence  0.00000  0.000000
Balanced Accuracy     0.50000  0.500000
```

Results and Discussion

Metrics

The metric for model-1 to model-2 shown in the table below. High R-squared suggests a good fit of the model-1 and model-2 to the data. While we know that model-1 only can predict 0.3% of the datapoint and accuracy is 0%. Similarly, model-2 did not make a good prediction for the evaluation set, which could be viewed in Figure 20.

Table 9. Metrics of Model-1 and Model-2

Model <fctr>	R_squared <fctr>	LogLike <fctr>	AIC <fctr>	variable <fctr>
Multiple Linear Regression	0.8893085	-62353.04	126322.1	53
Multiple Linear Regression - dummy encode predictor	0.8506650	-62861.80	125955.6	21

Table 10. Metrics of Model-3 to Model-6

Ordinal.Classification.Model <fctr>	Accuracy <fctr>	Error.Rate <fctr>	Kappa <fctr>
Ordinal Logistic Regression (OLR)	0.3698469	0.6301531	0.2670606
Random Forest	0.7726737	0.2273263	0.7361337
Generalized Linear Models with L1 Penalty	0.4958775	0.5041225	0.4079479
Classification Trees	0.6325088	0.3674912	0.5722134

Prediction accuracy

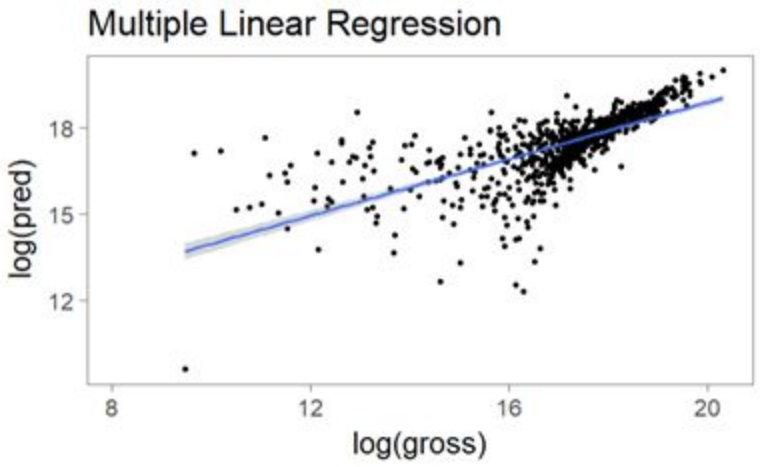


Figure 20. Predicted vs. actual gross in model-2 (Multiple Linear Regression) for the evaluation set with binary features.

Figure 21. Predicted vs. actual gross in OLC model for the evaluation set with binary features.

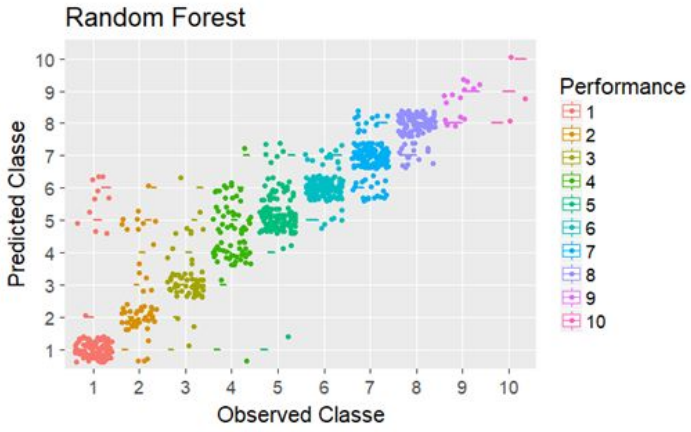


Figure 22. Predicted vs. actual gross in RF model for the evaluation set with binary features.

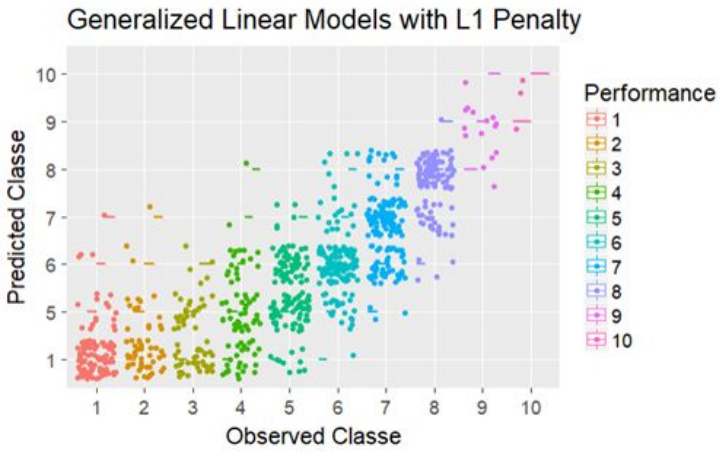


Figure 23. Predicted vs. actual gross in GLM model for the evaluation set with binary features..

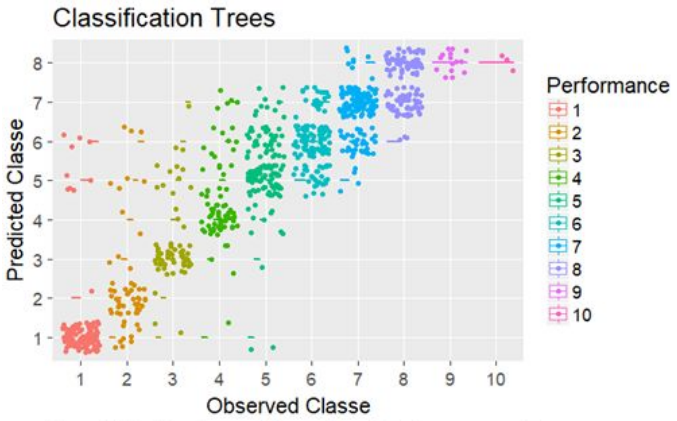


Figure 24. Predicted vs. actual gross in Classification Trees model for the evaluation set with binary features.

Conclusion and Discussions

The performance of the more complex Random Forest model and Classification Trees were better than that of simpler model Multiple Linear Regression and Generalized Logistic Regression (Table 10). Random Forest did a little better than on the evaluation set (77%) in comparison to Classification Trees (63%).

Generating new variables of the average box office made by the previous movies that the director or actors were involved had positively contributed to all of the models. Before using these variables, multicollinearity test was done to exclude any collinearity that could be brought by the newly created variables.

Comparing the multiple linear regression and the classification or ordinal regression, latter is a better approach to predict the gross on our dataset. Probably log transform will make the dependent variables normally distributed and fit the ordinal logistic regression and generalized linear regression models better.

Encode the categorical variable to dummy variable can reduce the complexity of computation. When using General Logistic Model with L1 penalty to fit the data, untransformed genres, content_rating and aspect_ratio dramatically slow down the computation. The dummy encoding improves the computation speed.

The sparsity of data in certain buckets, like class-10, is problematic in all the models(Table 8). The sensitivity similarly, The buckets with less sparsity, like class-2, class-3, class-4, class-9 are the vulnerable buckets whose accuracy is lower than other data-rich buckets. Similarly, the sparsity of data in independent variable will introduce meaningless values like NA or infinity during

computation. So some binary variable with only a few 1s had to be dropped in order to allow the computation to go through. One thing worth to be mentioned is that introducing 0 as one level of classes of dependent variable will be problematic because the ordered classes will be used as a number in the computation. Using 0~9 as buckets' labels will get a different model from using numbers, not including 0 as buckets' labels.

The movie release year was collected in the dataset but this variable only was chosen as a predictor in RF model. The month when the movie released is more interesting than the year because movie releasing in the month near a holiday like Christmas or summer will make more box office. In the future, we will collect the month information to see whether this information will be useful for improving the prediction model.

Reference

1. Box office revenue in the United States from 2012 to 2021 (in billion U.S. dollars) (https://www.statista.com/statistics/259988/box-office-revenue-in-the-us/)
2. Thomas W. Yee, The VGAM Package for Categorical Data Analysis.
3. Giuliano Galimberti, Gabriele Soritti, Matteo Di Maso Classification Trees for Ordinal Responses in R: The rpartScore Package.
4. Kellie J. Archer, glmpathcr: An R Package for Ordinal Response Prediction in High-dimensional Data Settings.
5. Benjamin Flora, Thomas Lampo, and Lili Yang, Predicting Movie Revenue from Pre-Release Data. Chance. 13:15-24 (2000).

Appendix

R code:

https://github.com/YunMai-SPS/DATA621_homework/blob/master/data621_final_project/DATA621_final_group4.Rmd