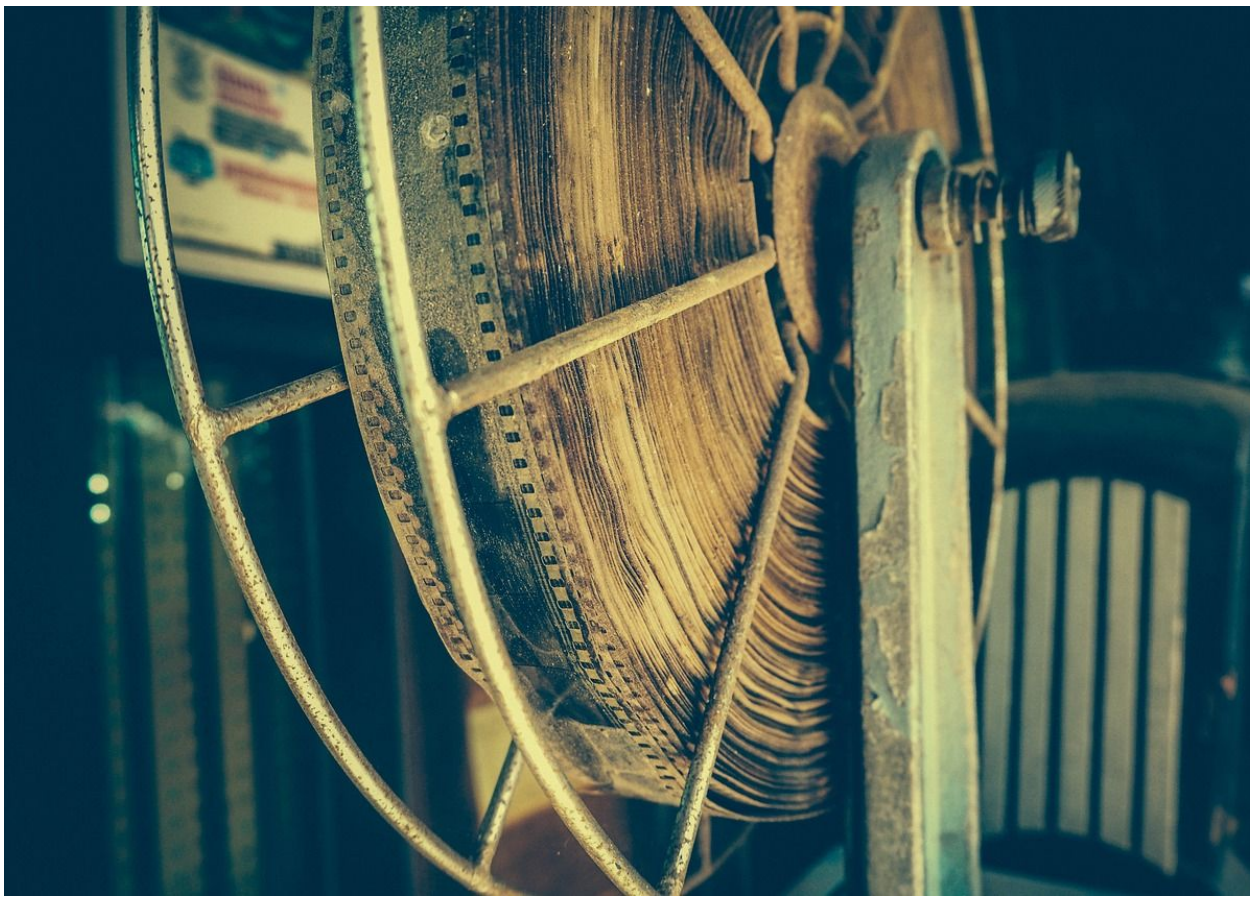
Business Analytics and Data Mining

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 endeavour towards performing exploratory data analysis on a movie related dataset. Our intentions behind this exploration is 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. Greater value of gross, greater will be the popularity of 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.

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**

Summary of the dataset

Highlights of main variables

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

Variable Structure

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.

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

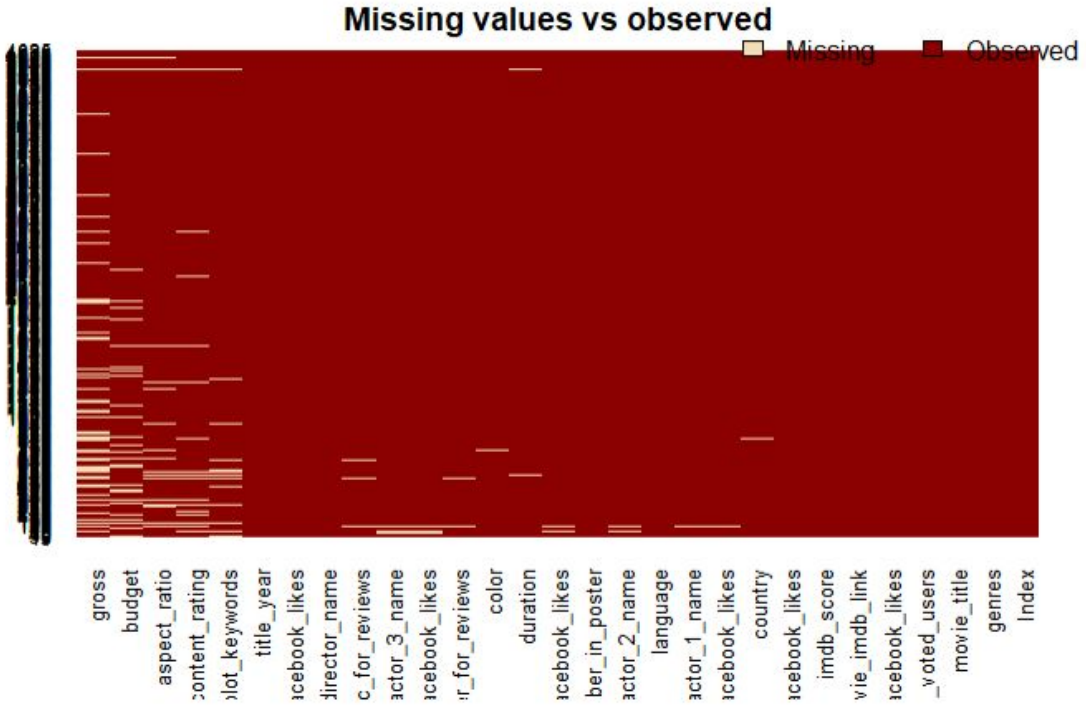
	se
num_critic_for_reviews	1.93
duration	0.39
director_facebook_likes	44.25
actor_3_facebook_likes	27.12
actor_1_facebook_likes	177.00
gross	1184144.49
num_voted_users	2218.79
cast_total_facebook_likes	239.89
facenumber_in_poster	0.03
num_user_for_reviews	6.04
budget	3725318.51
title_year	0.20
actor_2_facebook_likes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

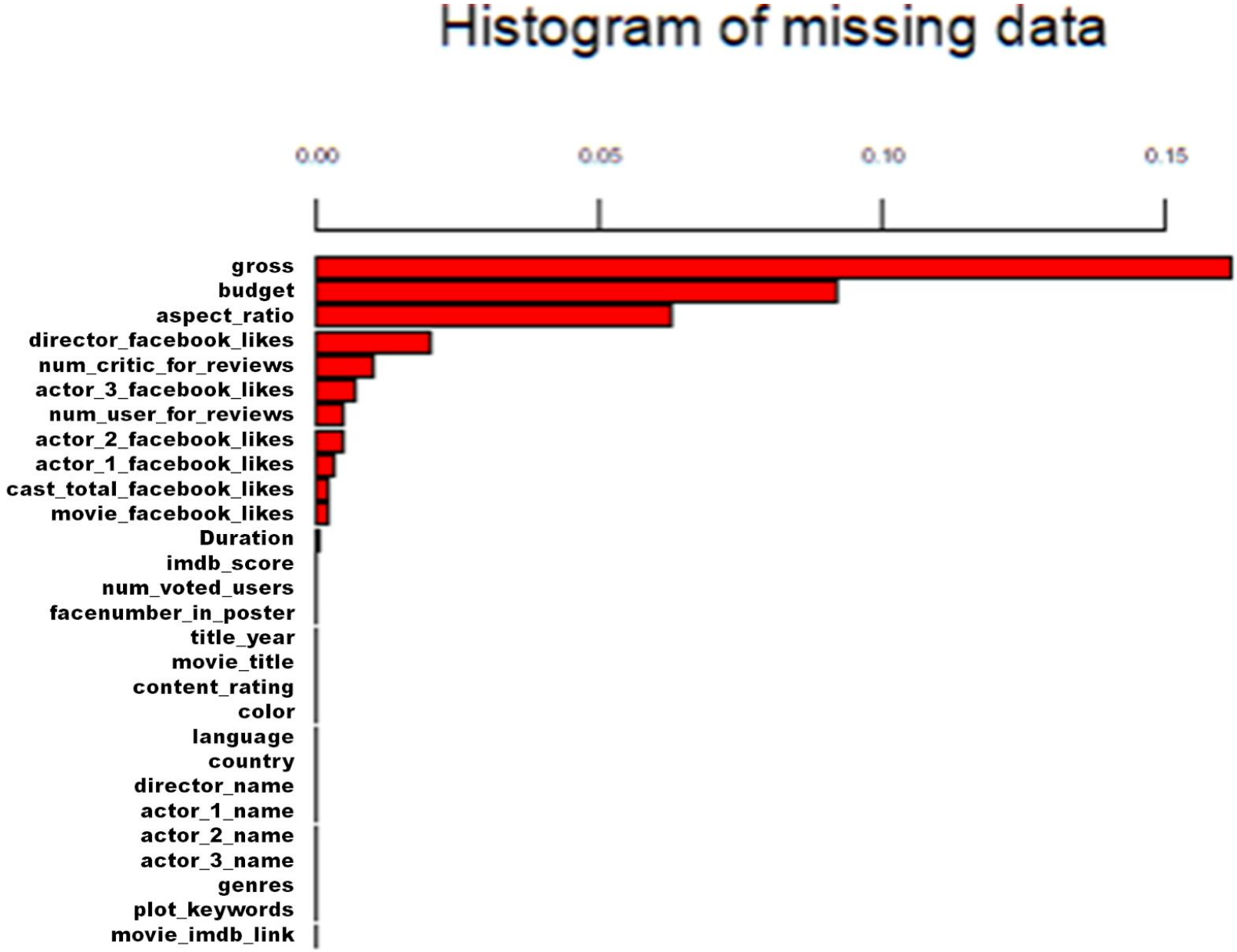
There are 1035 records have missing values and 476 records missed more than 5% data points. So let's see missing data in more detail.

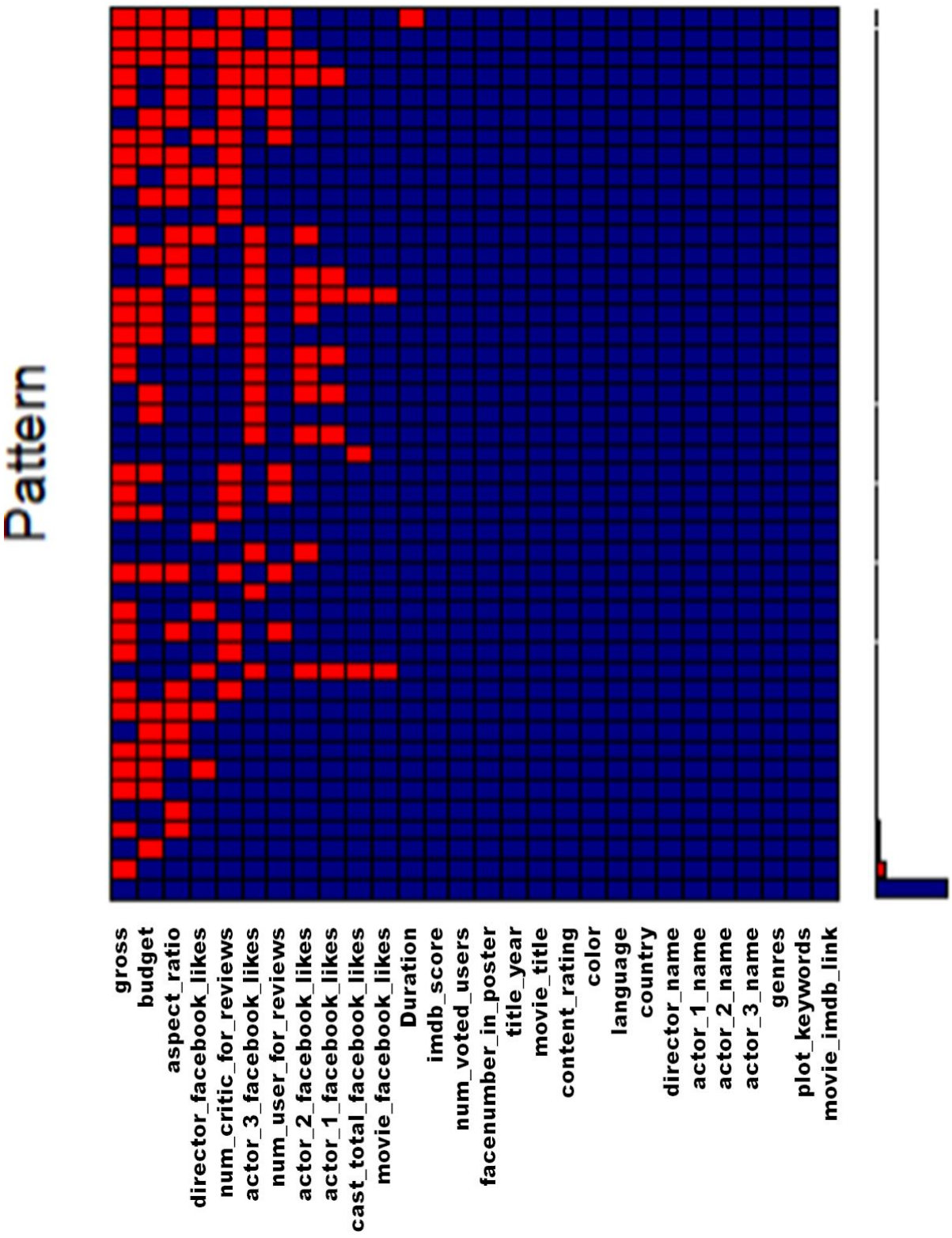
Missing Values

```
Variables sorted by number of missings:
                    Variable       Count
                       gross 0.175712515
                      budget 0.098389095
                aspect_ratio 0.065427509
              content_rating 0.060470880
               plot_keywords 0.029244114
                  title_year 0.021809170
               director_name 0.020817844
   director_facebook_likes 0.020817844
    num_critic_for_reviews 0.009665428
    actor_3_facebook_likes 0.004956629
              actor_3_name 0.004956629
      num_user_for_reviews 0.004213135
                     color 0.003965304
                  duration 0.002973978
              actor_2_name 0.002726146
        facenumber_in_poster 0.002726146
     actor_2_facebook_likes 0.002726146
                  language 0.002478315
     actor_1_facebook_likes 0.001486989
              actor_1_name 0.001486989
                   country 0.001239157
                    genres 0.000000000
               movie_title 0.000000000
           num_voted_users 0.000000000
 cast_total_facebook_likes 0.000000000
            movie_imdb_link 0.000000000
                imdb_score 0.000000000
      movie_facebook_likes 0.000000000
```

We noticed that there are many missing values for budget,aspect ratio and gross. Same is depicted in below missing values graphics and histogram.

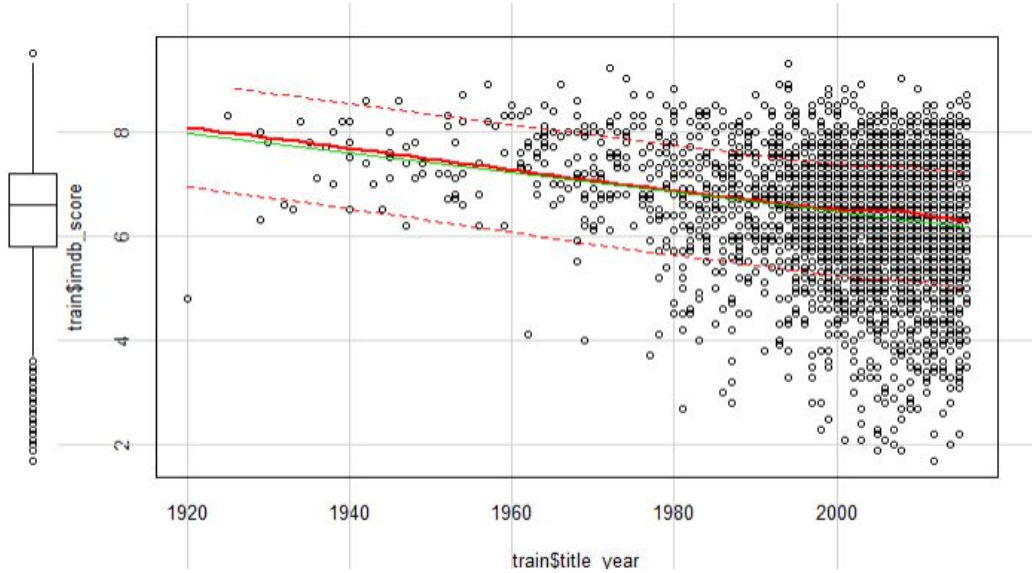



Histogram of missing data
0.00
0.05
0.10
0.15
gross
budget
aspect_ratio
director_facebook_likes
num_critic_for_reviews
actor_3_facebook_likes
num_user_for_reviews
actor_2_facebook_likes
actor_1_facebook_likes
cast_total_facebook_likes
movie_facebook_likes
Duration
imdb_score
num_voted_users
facenumber_in_poster
title_year
movie_title
content_rating
color
language
country
director_name
actor_1_name
actor_2_name
actor_3_name
genres
plot_keywords
movie_imdb_link

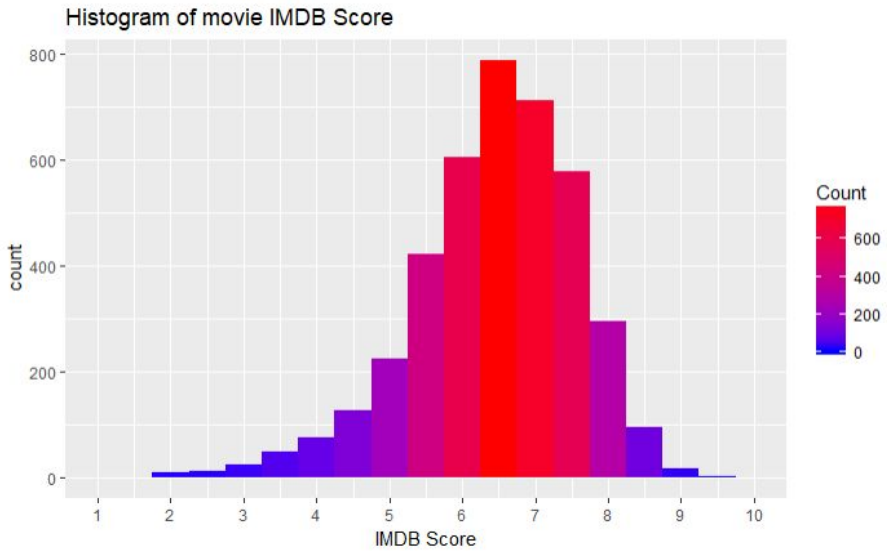


Visualization of title Year vs. Score

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.

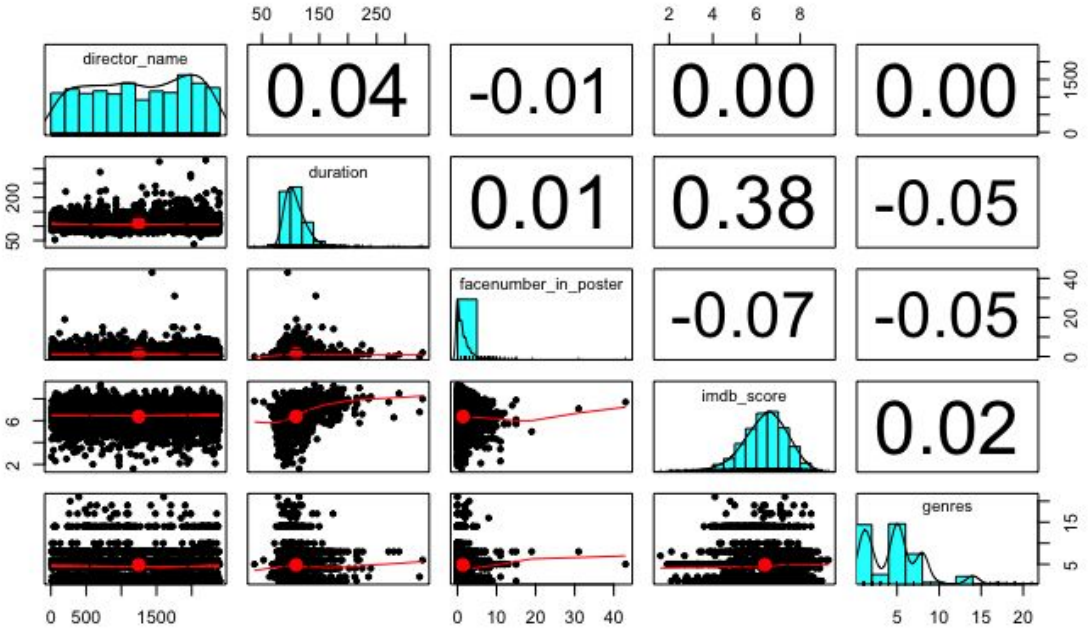


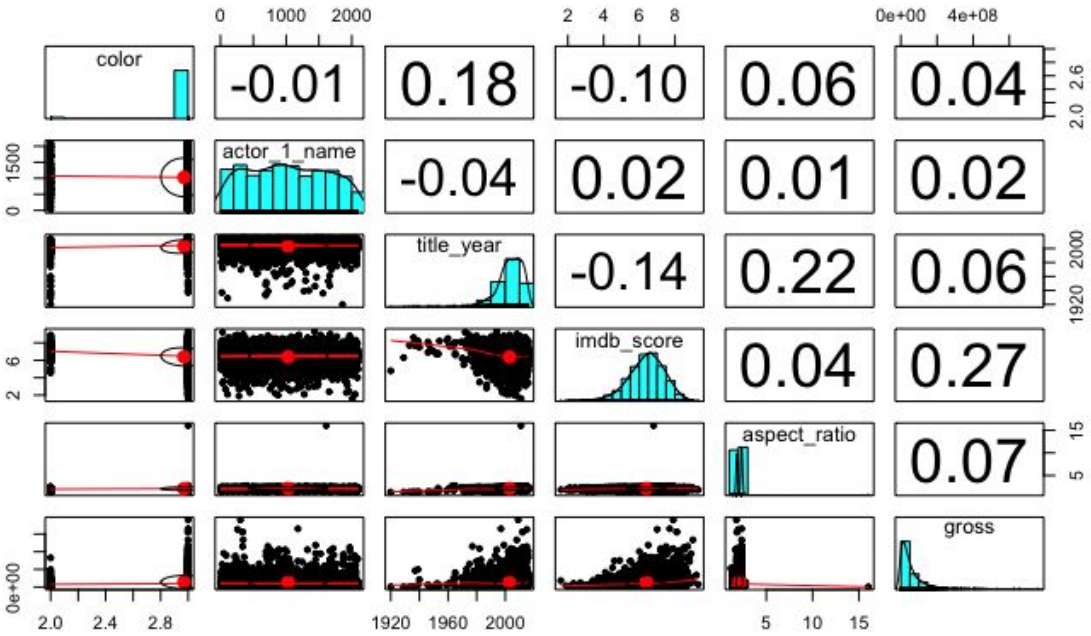
IMDB score looks normal.Many movies scored from 5 to 8 and few movies have score less than 4 or greater than 8.



Exploring correlation

As we can see from correlation plots only duration and IMDB score has a high correlation.



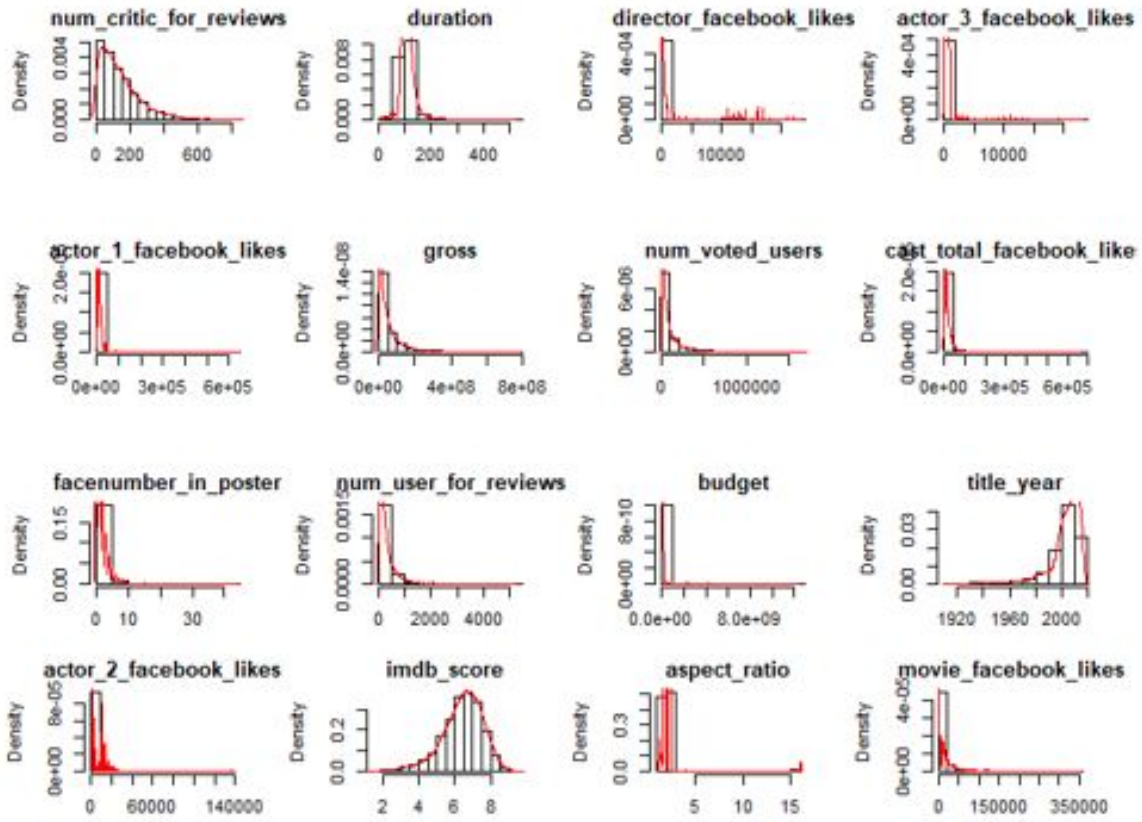


face number in posters has a negative correlation with IMDB score. *genre* has little correlation with score Interesting, director name has no correlation with IMDB score.Color and title year has highly positive correlation. Color and aspect ratio,gross has smaller positive correlations. Actor1_name has very small positive correlation with gross, meaning who plays the movies does not have impact on the gross.

Title year and aspect ratio and color are highly positively correlated. IMDB score has very small positive correlation with actor 1 name ,which means who was the actor 1 does not make the movie has a higher score. Interestingly, IMDB score has a negative correlation with title year,which means the old movies seems to have a higher score. IMDB and aspect ratio has small positive correlation.

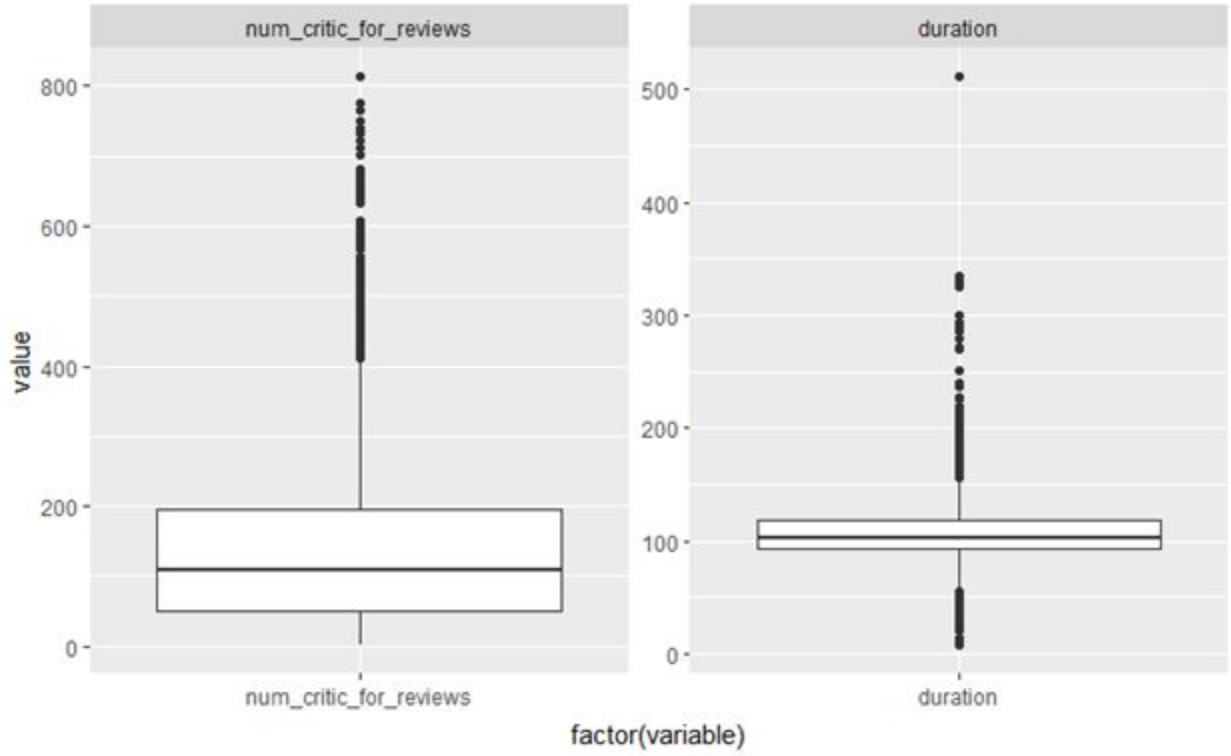
IMDB has a strong positive correlation with gross.

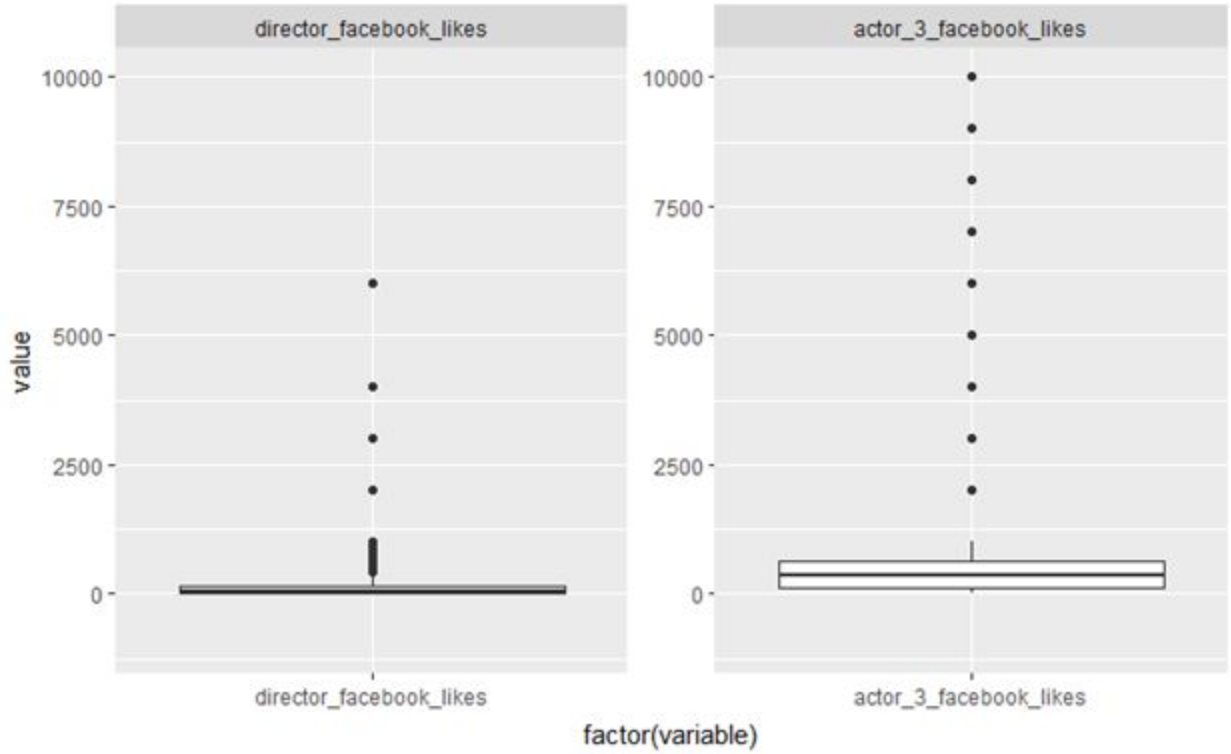
Correlation matrix

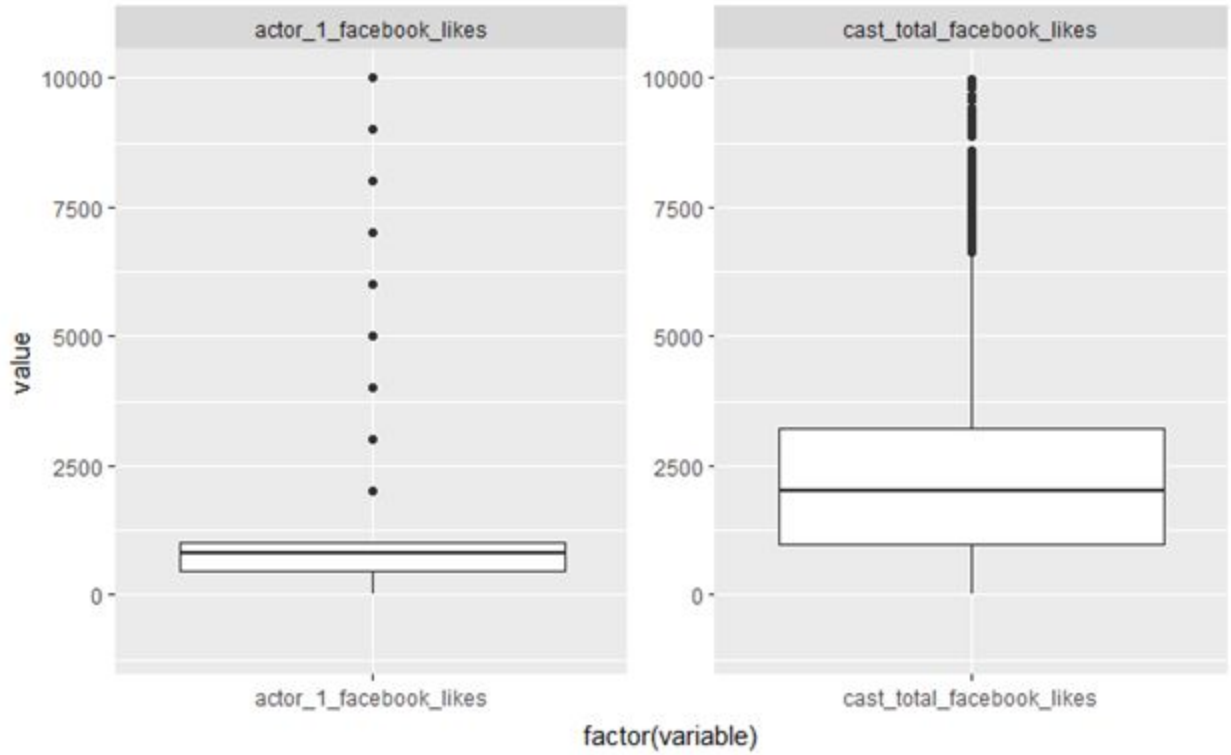


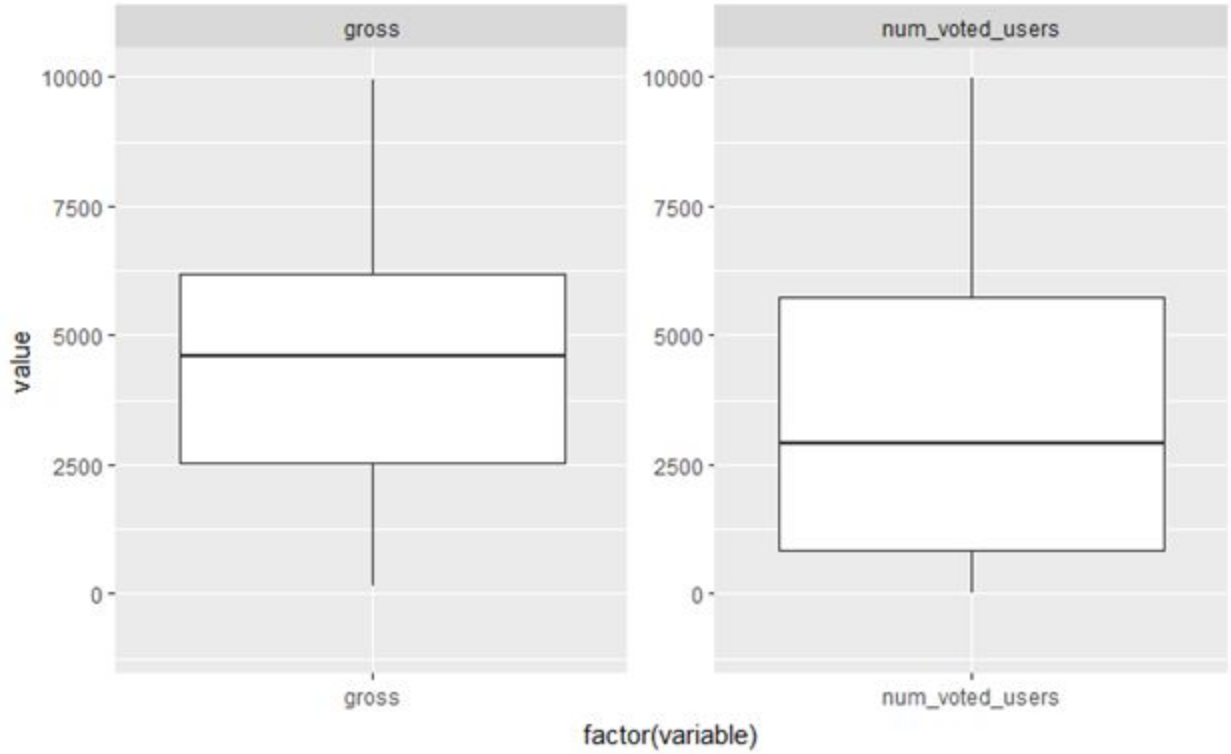
Outliers

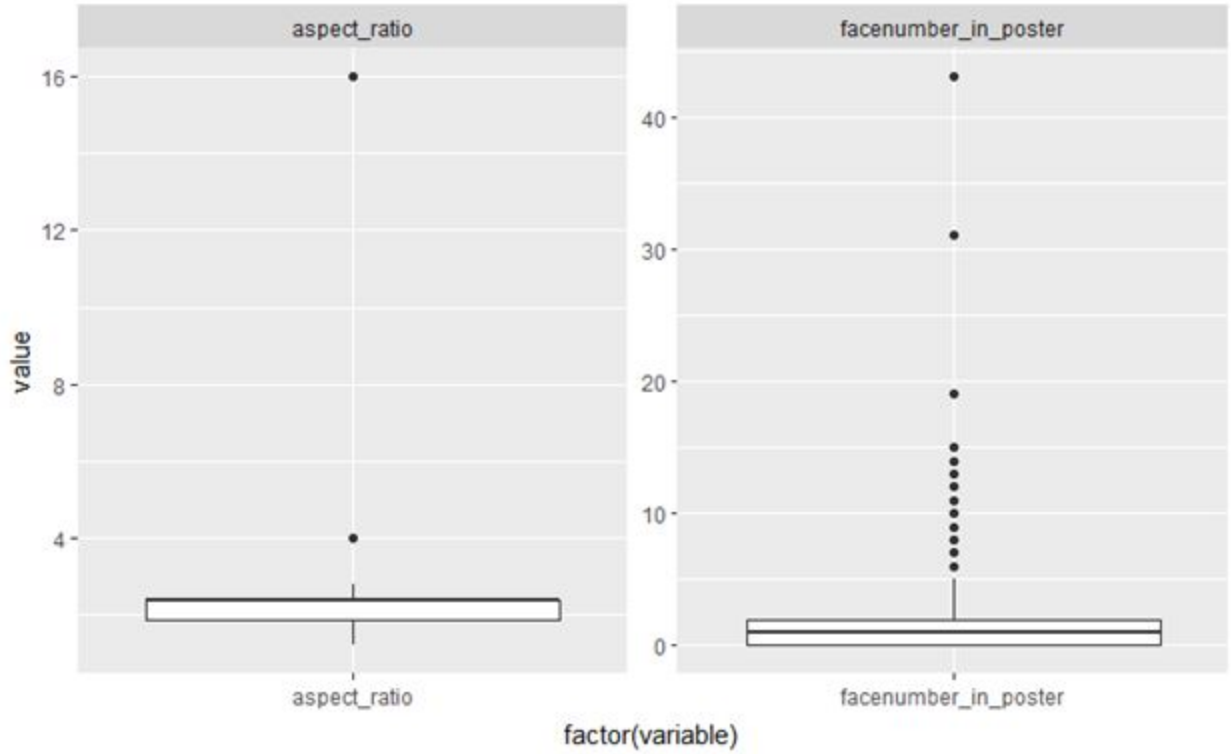
Box Plots

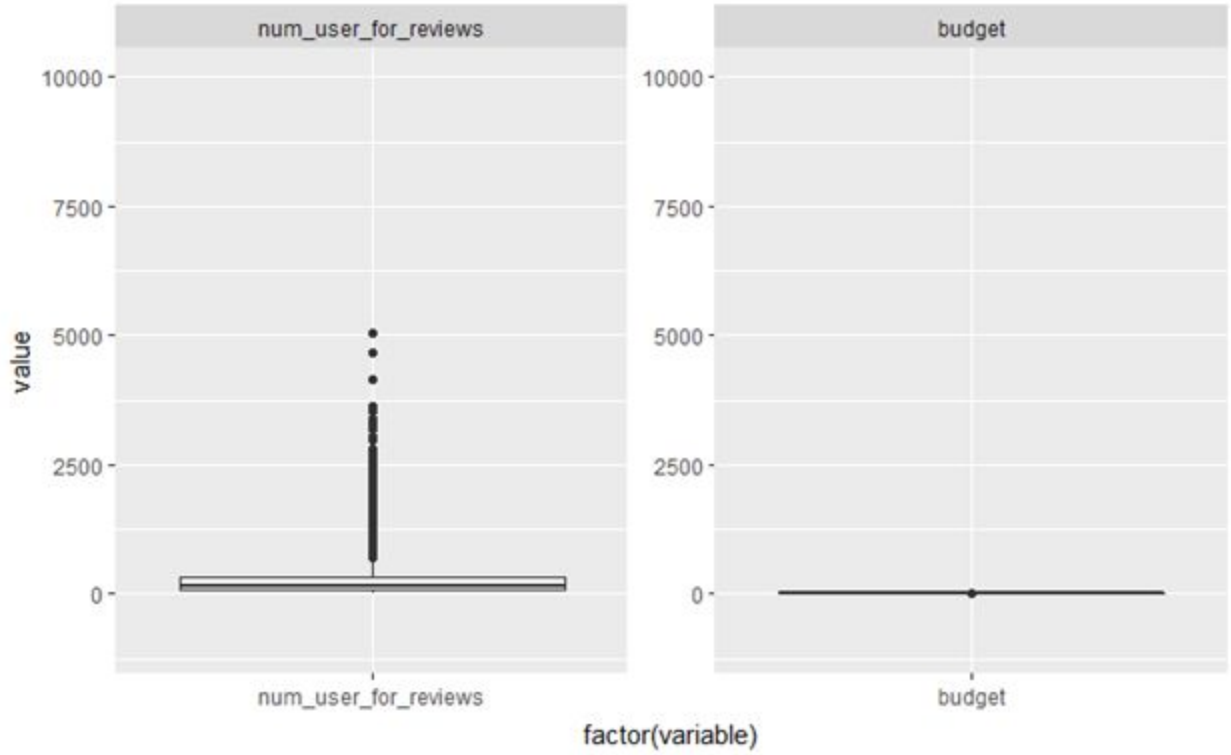


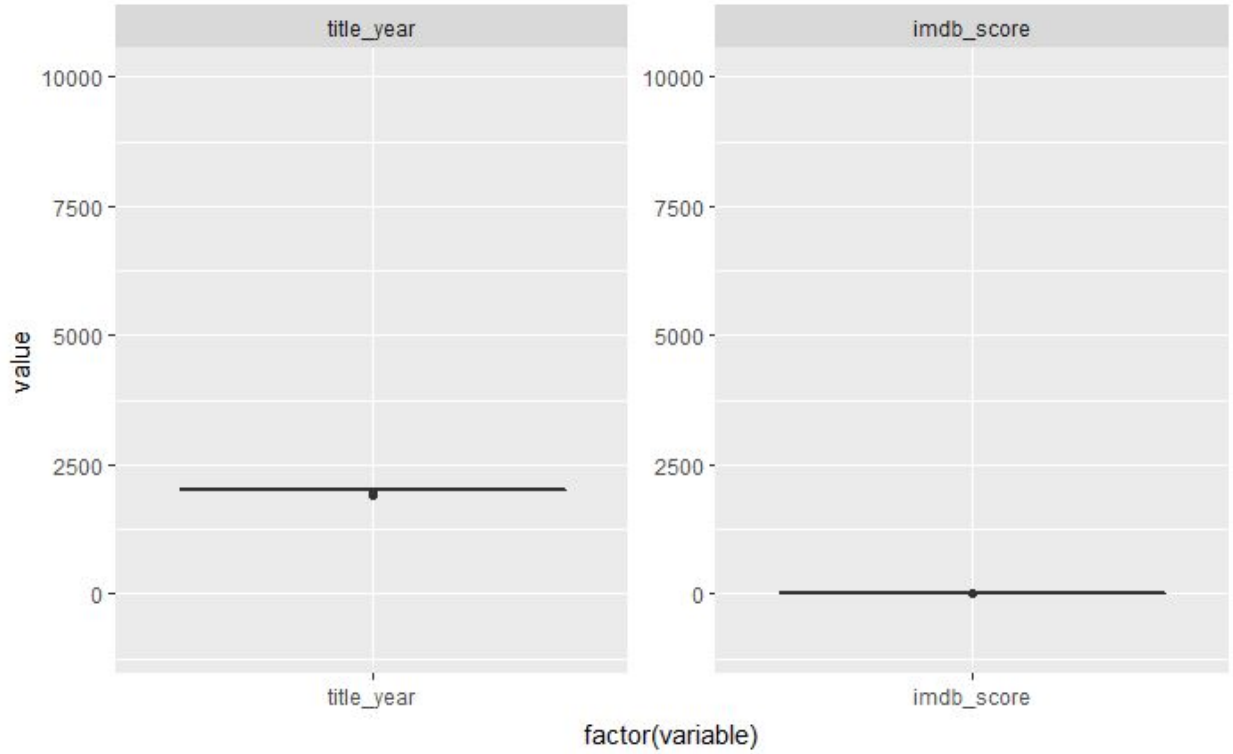

director_facebook_likes
actor_3_facebook_likes
10000
7500
5000
2500
0
value
director_facebook_likes
actor_3_facebook_likes
factor(variable)

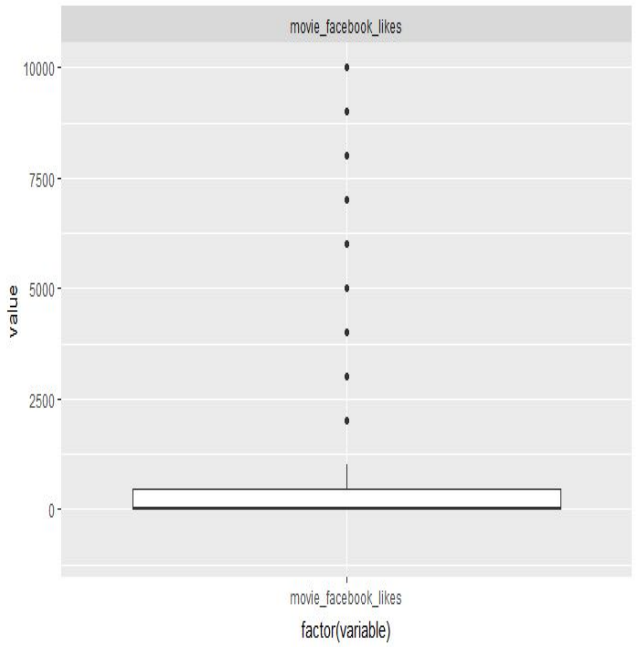

actor_1_facebook_likes
cast_total_facebook_likes
10000
7500
5000
2500
0
value
actor_1_facebook_likes
cast_total_facebook_likes
factor(variable)

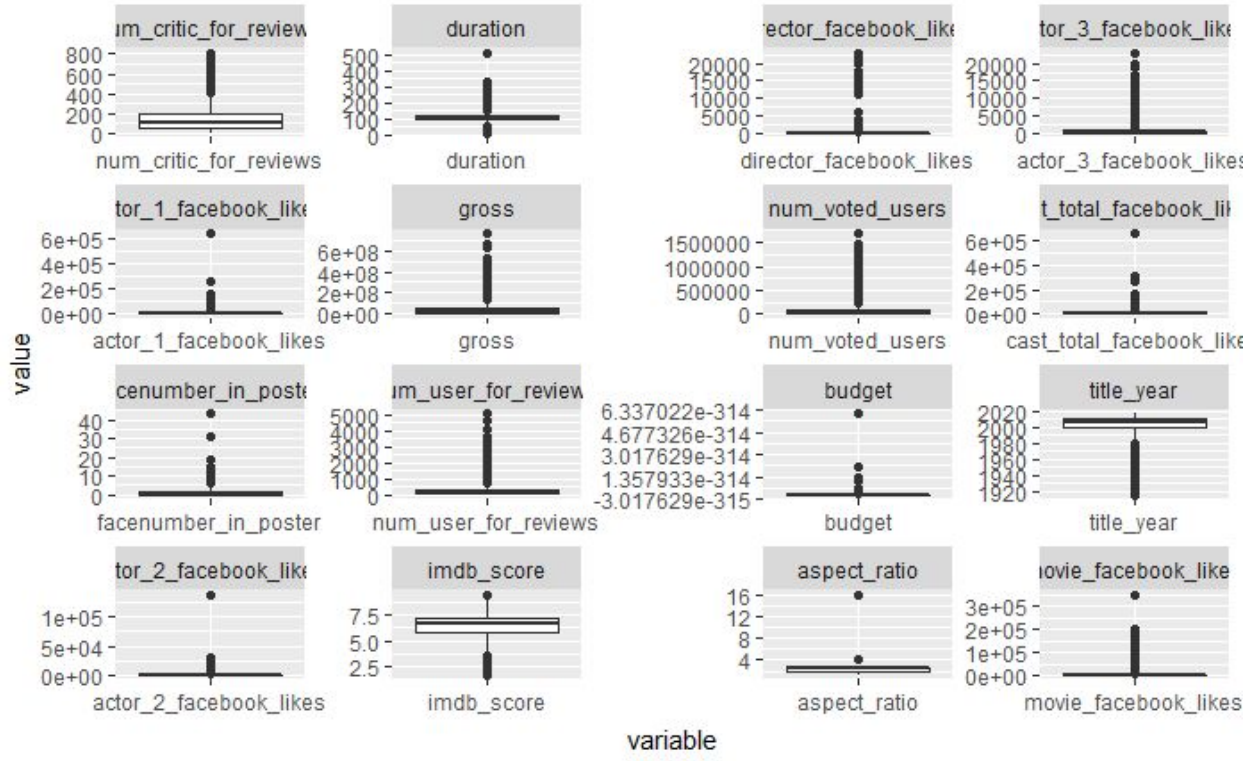

gross
num_voted_users
10000
7500
5000
2500
0
value
gross
num_voted_users
factor(variable)


aspect_ratio
facenumber_in_poster
16
12
8
4
40
30
20
10
0
value
aspect_ratio
facenumber_in_poster
factor(variable)


num_user_for_reviews
budget
10000
7500
5000
2500
0
value
num_user_for_reviews
budget
factor(variable)


title_year
imdb_score
10000
7500
5000
2500
0
value
title_year
imdb_score
factor(variable)


movie_facebook_likes
10000
7500
5000
2500
0
value
movie_facebook_likes
factor(variable)



Data Preparation

Missing Values: We will check percentage of missing data in dataset. .

Variable color has the missing values, general rule of thumb should be to set a year as a threshold for color movies, for example there were only few black and white movies after 1990. After 1990 there were 125 black and white movies, 4,214 color movies and 14 missing values. We will replace these values with color and the missing values before 1990 with black and white.

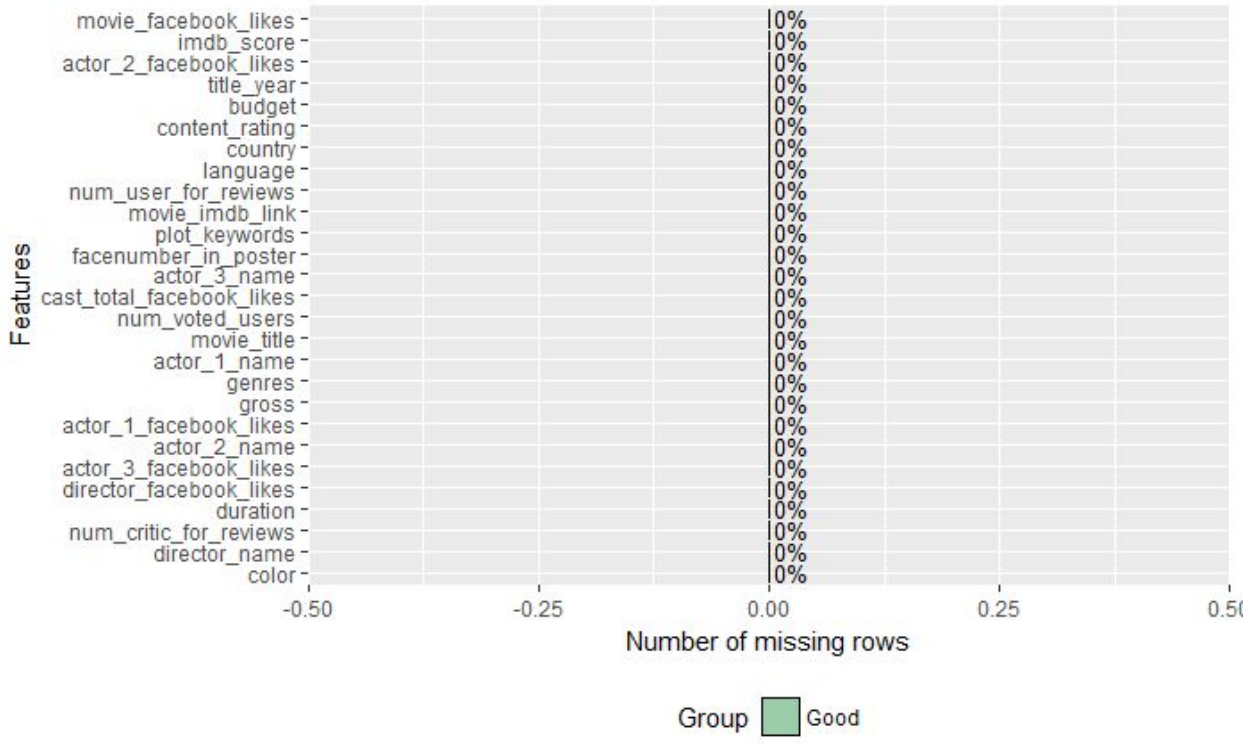
Director Name – If there is no value in director's name or NA, we will replace with None. For all these cases we will replace director Facebook likes with zero.

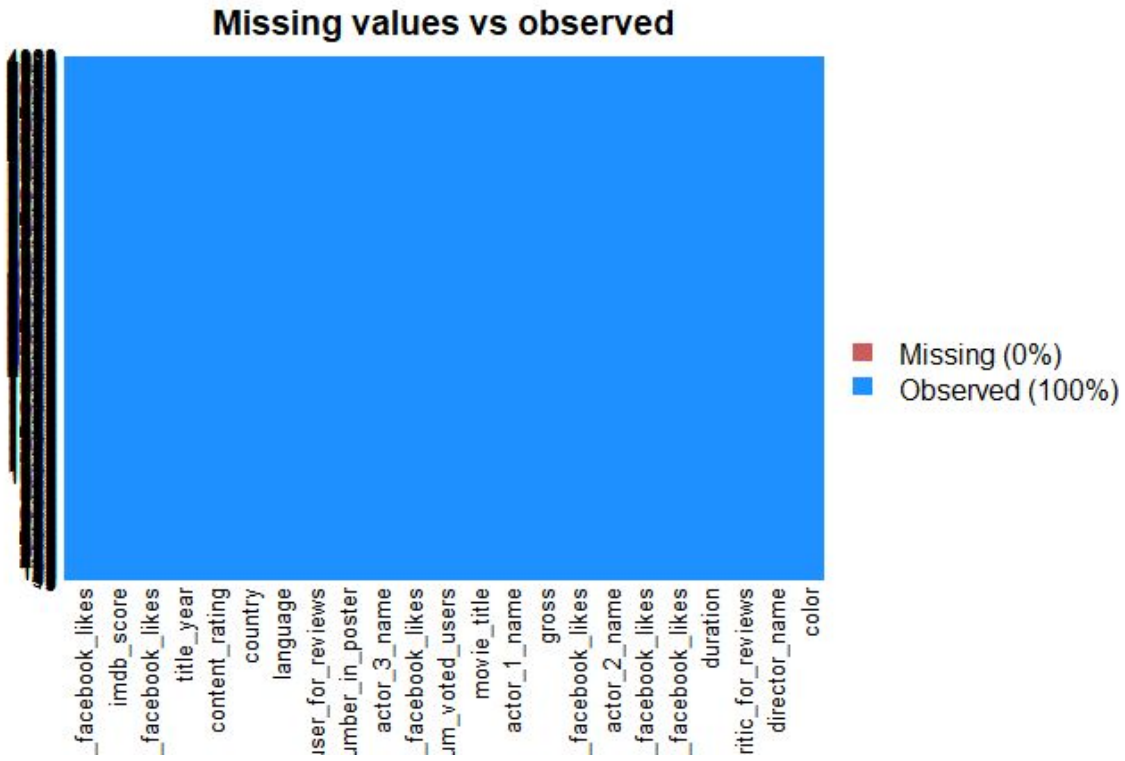
Actor Facebook likes- Imputation with mean or median is not a better approach in replacing missing values. In that case, we will replace the missing values with zero. All three variables

for actor Facebook likes and movie Facebook likes, cast total Facebook likes will be replaced with zero if there are missing values.

Variable Gross has 18% missing values in the dataset. Removing the variable from analysis might be a better choice. Since this variable is correlated with imdb score, removing this variable from analysis can be a risk of losing information. We will try to treat missing values for this variable. We are making the assumption that higher the imdb score, higher will be the gross and more popular movie will be. We will get the median and mean imdb score for the records with missing values, the median was 6.6 and mean was 6.4 for these values. Second step was to get median value of variable gross with median imdb score of 6.6. The median gross value for this category was 25,517,500. We will impute the missing values for gross with these values.

Variable aspect ratio will be removed from the analysis. Missing values in the variable duration will be replaced by median value of duration.





Outliers: There are a lot of outliers in the dataset. 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 loss of information. Treatment of outliers might not produce accurate results as the imputation and score. For comparison purposes, we will create another variables in the outlier treatment process to run additional model based on these newly created variables.

Normality: Although it is suggested to check the normality of variables prior to setting up the models, we will set up non parametric regression models to overcome this issue, if exists.

Multicollinearity: Multicollinearity check is another step in the data mining process. In order to remove the variables for redundant information in the dataset, we will compare

the correlation coefficient between the variables and check for the independent variables having high correlation between them. Correlation between variables was calculated. The pair of variables with greater than correlation coefficient of 0.5 were selected and marked as dependent or multicollinear. The variables from these pair were selected in order to remove singularities.

Variable 1	Variable 2	value
cast_total_facebook_likes	actor_1_facebook_likes	0.95
actor_1_facebook_likes	cast_total_facebook_likes	0.95
num_user_for_reviews	num_voted_users	0.80
num_voted_users	num_user_for_reviews	0.80
movie_facebook_likes	num_critic_for_reviews	0.69
num_critic_for_reviews	movie_facebook_likes	0.69
num_voted_users	gross	0.64
gross	num_voted_users	0.64
actor_2_facebook_likes	cast_total_facebook_likes	0.63
cast_total_facebook_likes	actor_2_facebook_likes	0.63
num_voted_users	num_critic_for_reviews	0.62
num_critic_for_reviews	num_voted_users	0.62
num_user_for_reviews	num_critic_for_reviews	0.61
num_critic_for_reviews	num_user_for_reviews	0.61
num_user_for_reviews	Gross	0.56
gross	num_user_for_reviews	0.56
actor_2_facebook_likes	actor_3_facebook_likes	0.56
actor_3_facebook_likes	actor_2_facebook_likes	0.56
movie_facebook_likes	num_voted_users	0.53
num_voted_users	movie_facebook_likes	0.53

Build Models

Model-1: Multiple Linear Regression I

	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|Family|Fantasy	6.390156e+07	1.712222e+07	3.732083	1.939859e-04
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

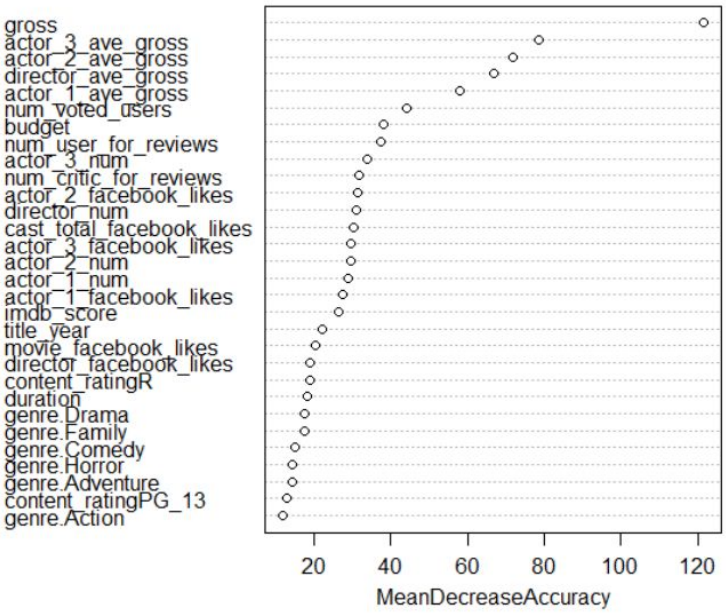
Model-2: Multiple Linear Regression II

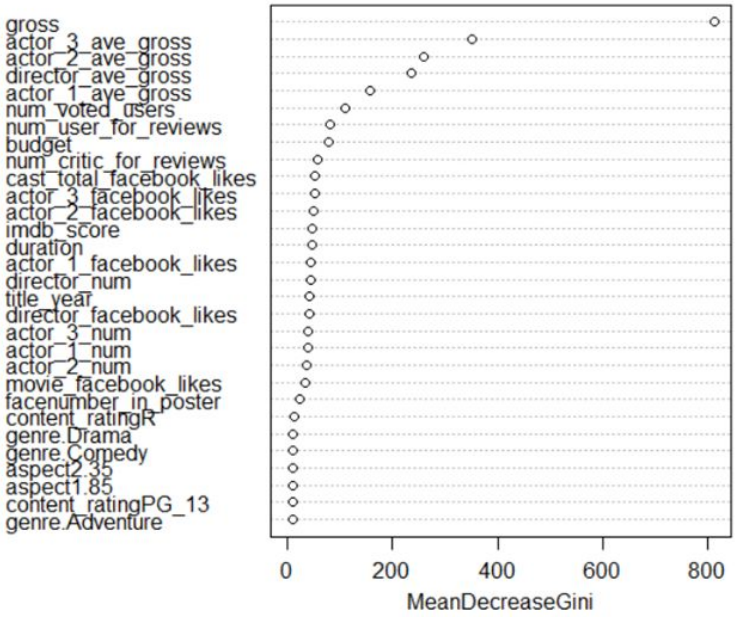
	Estimate <dbl>	Std..Error <dbl>	t.value <dbl>	Pr...t.. <dbl>
(Intercept)	2.742317e+08	1.382791e+08	1.983176	4.743100e-02
num_voted_users	7.220259e+01	6.729021e+00	10.730029	2.007113e-26
budget	4.380149e-02	8.841787e-03	4.953918	7.640960e-07
title_year	-1.337406e+05	6.683453e+04	-2.001071	4.546697e-02
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

Model-3: Ordinal Logistic Regression (OLR)

```
Confusion Matrix and Statistics

          Reference
Prediction   1  10   2   3   4   5   6   7   8   9
        1  412   0   0 136 128 119 131  99  31   0
        10   0   0   0   0   0   0   0   0   0   0
        2    3   0   0  10   2   3   2   2   0   0
        3    4   0   0   1   7   4   2   8   0   0
        4    2   0   0   3   3   6   2   3   1   0
        5   18   0   0  21  25  29  77  35  12   0
        6   37   0   0  30  54  96 211 391 155   5
        7    1   0   0   3   6  15  22  80 126  15
        8    0   1   0   1   5   6  11  35 162 140
        9    2  42   3   1   4   5  12  25  80 275

Overall Statistics

               Accuracy : 0.3452
                 95% CI : (0.3292, 0.3615)
    No Information Rate : 0.1995
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.2384
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 10  Class: 2  Class: 3  Class: 4 Class: 5 Class: 6  Class: 7 Class: 8 Class: 9
Sensitivity            0.8601   0.00000 0.0000000 0.0048544 0.0128205 0.102473   0.4489   0.11799  0.28571  0.63218
Specificity            0.7794   1.00000 0.9935199 0.9921679 0.9946271 0.939647   0.7377   0.93088  0.92971  0.94128
Pos Pred Value         0.3902       NaN 0.0000000 0.0384615 0.1500000 0.133641   0.2155   0.29851  0.44875  0.61247
Neg Pred Value         0.9714   0.98735 0.9991114 0.9392052 0.9316163 0.920151   0.8929   0.80895  0.86664  0.94574
Prevalence             0.1410   0.01265 0.0008829 0.0606239 0.0688640 0.083284   0.1383   0.19953  0.16686  0.12802
Detection Rate         0.1212   0.00000 0.0000000 0.0002943 0.0008829 0.008534   0.0621   0.02354  0.04768  0.08093
Detection Prevalence   0.3108   0.00000 0.0064744 0.0076516 0.0058858 0.063861   0.2881   0.07887  0.10624  0.13214
Balanced Accuracy      0.8198   0.50000 0.4967599 0.4985111 0.5037238 0.521060   0.5933   0.52444  0.60771  0.78673
```

Model-4: Random Forest


gross
actor_3_ave_gross
actor_2_ave_gross
director_ave_gross
actor_1_ave_gross
num_voted_users
budget
num_user_for_reviews
actor_3_num
num_critic_for_reviews
actor_2_facebook_likes
director_num
cast_total_facebook_likes
actor_3_facebook_likes
actor_2_num
actor_1_num
actor_1_facebook_likes
imdb_score
title_year
movie_facebook_likes
director_facebook_likes
content_ratingR
duration
genre.Drama
genre.Family
genre.Comedy
genre.Horror
genre.Adventure
content_ratingPG_13
genre.Action
20
40
60
80
100
120
MeanDecreaseAccuracy


gross
actor_3_ave_gross
actor_2_ave_gross
director_ave_gross
actor_1_ave_gross
num_voted_users
num_user_for_reviews
budget
num_critic_for_reviews
cast_total_facebook_likes
actor_3_facebook_likes
actor_2_facebook_likes
imdb_score
duration
actor_1_facebook_likes
director_num
title_year
director_facebook_likes
actor_3_num
actor_1_num
actor_2_num
movie_facebook_likes
facenumber_in_poster
content_ratingR
genre.Drama
genre.Comedy
aspect2.35
aspect1.85
content_ratingPG_13
genre.Adventure
0
200
400
600
800
MeanDecreaseGini

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6   7   8   9  10
        1   64   0  23  43  33  61  86  76  49   5
        2    0   0   0   0   0   1   2   1   0   0
        3   31   0  10  16  16  27  35  42  23   5
        4   29   0  18   9  18  41  49  42  31   3
        5   39   0  16  17  21  32  52  42  35   2
        6   76   1  50  39  50  80 116  96  86   7
        7   79   0  30  35  56  82 106 100  77   7
        8   86   1  31  42  51  80 116  82  76   7
        9   70   0  24  25  32  60 101  81  52   7
        10   5   1   4   8   6   6  15   5   6   0

Overall Statistics

               Accuracy : 0.1248
                 95% CI : (0.1138, 0.1364)
    No Information Rate : 0.1995
    P-Value [Acc > NIR] : 1

                  Kappa : -0.0135
 Mcnemar's Test P-Value : NA


                     Class: 1 Class: 10  Class: 2  Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity            0.8580   0.00000 0.0000000 0.0048544 0.017094 0.102473  0.43617
Specificity            0.7756   1.00000 0.9932253 0.9924812 0.996839 0.939326  0.73497
Pos Pred Value         0.3856       NaN 0.0000000 0.0400000 0.285714 0.133028  0.20897
Neg Pred Value         0.9708   0.98735 0.9991111 0.9392232 0.932033 0.920126  0.89036
Prevalence             0.1410   0.01265 0.0008829 0.0606239 0.068864 0.083284  0.13832
Detection Rate         0.1210   0.00000 0.0000000 0.0002943 0.001177 0.008534  0.06033
Detection Prevalence   0.3137   0.00000 0.0067687 0.0073573 0.004120 0.064155  0.28870
Balanced Accuracy      0.8168   0.50000 0.4966127 0.4986678 0.506967 0.520900  0.58557
                     Class: 7 Class: 8 Class: 9
Sensitivity           0.11504  0.28395  0.63218
Specificity           0.93199  0.92971  0.94161
Pos Pred Value        0.29658  0.44722  0.61384
Neg Pred Value        0.80861  0.86636  0.94576
Prevalence            0.19953  0.16686  0.12802
Detection Rate        0.02295  0.04738  0.08093
Detection Prevalence  0.07740  0.10594  0.13184
Balanced Accuracy     0.52351  0.60683  0.78690
```

Appendix

R code:

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