MGSC 661 Midterm Project: The 2024 Prediction Challenge

Wicked 6

Arial Huang
Atharva Vyas
Liliana Garcia
Maggie Huang
Maral Batnasan
Thao Nguyen

McGill University
Desautels Faculty of Management
Master of Management in Analytics

October 24, 2024

1. Introduction

In this project, our team aims to develop a predictive model to estimate IMDb scores for 12 upcoming movies scheduled for release in October and November 2024. We employed both linear and non-linear regression models to capture the potential relationships between movie features and their corresponding ratings. To ensure robust model performance, we systematically addressed key regression challenges such as linearity violations, multicollinearity, outliers, and heteroscedasticity.

Our approach involves comprehensive data exploration, feature engineering, hyperparameter tuning, and model selection. Multiple models were trained using cross-validation and evaluated using various performance metrics. After thorough comparison and analysis, the most optimal model is selected for predicting the IMDb scores. The detailed methodology and findings will be elaborated in the following sections.

2. Data Description

Numerical Variables

Prior to analysis, we performed both visual tests (with qq plots) and Bonferroni tests to identify the outliers and removed 7 rows from the dataset (191, 316, 395, 492, 989, 1581, 1806).

First, we examined the dependent variable imdb_score. The IMDb scores are approximately normally distributed and slightly left-skewed. Most movies receive scores in the 5 to 7 range, with the median score at 6.60 (Appendix 1).

We then examined the relationship between the numeric variables and IMDb score (Appendix 2). The correlation analysis revealed several key relationships. The movie duration and the number of articles in the news of the main country about the film (nb_news_articles) show moderate positive correlations with IMDb score, indicating that longer films and news coverage tend to score higher. Conversely, the year of release (release_year), movie budget (movie_budget), number of faces on the main poster (nb_faces), and movie rankings from IMDbPro (movie_meter_IMDBpro) show negative correlations, suggesting that newer, higher budget movies with more people on the poster may not score better. Other variables have little to no correlation with IMDb score. In the following section, we will explore each numerical predictor in more detail:

a. movie_budget: The distribution is right-skewed, with 75% of the movies with budgets lower than 20 million dollars. There is a 99.95% probability that there is a statistically significant relationship between movie_budget and imdb_score.
b. release_year: The distribution shows an increasing trend of movie releases over time, especially after the 1980s. There is also a higher concentration of movies in recent years. This may reflect changes in movie production, technological advancements and availability, and viewer preferences

over time. The probability that there is a statistically significant relationship between release_year and imdb_score is almost 100%.

c. duration: Most movies' durations fall between 100 and 150 minutes. The scatter plot suggests that heteroskedasticity exists and duration has some potential association with IMDB scores, with shorter movies tending to have a lower IMDb score. The probability that there is a statistically significant relationship between release_year and imdb_score is almost 100%.
d. nb_news_articles: The distribution is right-tail heavy, which means that most movies have a small number of articles and only a few movies experienced a significant number of media coverage. From the scatterplot, there is also visual evidence that heteroskedasticity exists. In addition, there is almost a 100% chance that there is a statistically significant relationship between release_year and imdb_score.
e. actor1_star_meter (ranking of main actor), actor_2_star_meter (ranking of second main actor), and actor3_star_meter (ranking of third main actor): Most actors fall in the top 1000 rankings. All 3 variables do not have a statistically significant relationship with imdb_score.
f. nb_faces: The distribution is right-skewed, with most movies having under 5 faces appearing in the main poster. There is 99.99% chance that there is a statistically significant relationship between number of faces and imdb_score.
g. movie_meter_IMDBpro (popularity ranking): The distribution is right-skewed, with visual evidence that heteroskedasticity exists. There is 99.99% chance that there is a statistically significant relationship between movie_meter_IMDBpro and imdb_score.

In conclusion, other than actor's popularity, all numeric variables have a statistically significant relationship with imdb_score. *(Please refer to Appendix 3 for the histograms and scatter plots for each numerical variable).*

Categorial Variables

The bar charts for all the categorical variables can be found in Appendix 8. For director, distributer, cinematographer, production_company, and plot_keywords (variables i. – m.), only the top 30 categories are displayed, since there are too many categories.

The following variables are highly skewed towards one or more categories: country, language, maturity rating, colour_film, aspect_ratio. The imbalance suggests that any analysis involving these variables will likely be influenced heavily by the top categories due to their disproportionately large counts.

a. release_month: This variable is fairly evenly distributed across all 12 months. October has the highest number of movie releases, with 216, while May has the fewest, with 94 releases.
b. release_day: This variable is fairly evenly distributed across all 30 days within a month. The 25th has the highest number of movie releases, with 110, while the 30th has the fewest, with 44 releases.

c. country: This variable is highly skewed towards USA with an overwhelmingly dominant count of 1555. The second most frequent country, the UK, has significantly fewer occurrences, at 177, with a steep drop-off for all other countries.
d. language: This variable is highly skewed towards English with an overwhelmingly dominant count of 1892. The remaining languages have negligible representation. This makes sense, as most movies in this dataset are from English-speaking countries like USA and UK.
e. maturity_rating (The content rating of the film): This variable is highly skewed toward certain categories, with the "R" rating (1013 counts) being the most common, followed by "PG-13" (582 counts) and "PG" (255 counts). Other categories show much lower representation.
f. colour_film: This variable is highly skewed toward color, with a count of 1867. Only 63 films are in black and white. This distribution reflects the general trend in the film industry of producing an increasing number of color films over the years.
g. aspect_ratio (The aspect ratio of the image of the film): this variable is dominated by two aspect ratios: 2.35 (with 981 counts) and 1.85 (with 853 counts). This suggests that the dataset primarily consists of films using these standard aspect ratios, which are common in cinematic releases.
h. genres: this variable is relatively skewed. The genre "drama" is by far the most common, with 1060 occurrences, followed by "thriller" (575 counts), "romance" (473 counts), "crime" (417 counts), and "action" (387 counts). Other genres have fewer representations.
i. director: Woody Allen is the most represented director with 18 films. After him, the number of films by other directors gradually decreases, ranging from 12 to 1.
j. distributor: Warner Bros is the most represented distributor with 169 films, followed by Universal Pictures (146) and Paramount Pictures (138). After the top few distributors, the number of films by other distributors decreases gradually.
k. cinematographer: There is a significant category labeled as "multiple" with 79 counts, indicating that many films may involve collaborations between cinematographers or a mix of credits. Roger Deakins is the most represented individual cinematographer with 18 films, and the number of films by other individual cinematographer decreases gradually.
l. production_company: Universal Pictures is the most represented producer, with 110 films, followed closely by Paramount Pictures (99) and Warner Bros. (96). After the top 6 producers, the number of films by other producers decreases gradually.
m. plot_keywords (keywords in the main IMDb plot of the movie): The most frequent keyword is "murder," with 84 occurrences, followed by a gradual decline in the counts for other keywords.

3. Model Selection

In the exploratory phase, we addressed some fundamental concerns in regression modeling: non-linearity, heteroskedasticity, outliers, and collinearity. Seven observations identified as outliers were removed. We also detected non-linearity and heteroskedasticity in certain predictors, though no evidence of collinearity was found (see Appendix 4 for details). After resolving these issues, a total of 15 predictors were retained in the final model, reflecting the processed characteristics of the data.

Numerical Predictors

The selected numerical predictors include movie_budget, duration, release_year, nb_news_articles, nb_faces, movie_meter_IMDBpro (popularity ranking), and min_star_meter (most popular actor ranking). The following adjustments were made:

a. Log Transformation

   To address skewness and non-linearity, log transformations were applied to several variables. This method proved effective for specific predictors. As demonstrated in the Appendix 5, movie_budget, duration, and actor3_star_meter exhibited improved linearity post-transformation. Additionally, log transformations mitigated skewness in duration, nb_news_articles, actor1_star_meter, actor2_star_meter, actor3_star_meter, and movie_meter_IMDBpro. Consequently, subsequent analysis and modeling incorporated these predictors in their log-transformed form.

b. Correcting Heteroskedasticity

   The presence of heteroskedasticity was confirmed through the Non-constant Variance (NCV) test ($p = 0.2967$). To address this issue, heteroskedasticity-robust standard errors were employed in the regression analysis. Although correcting for heteroskedasticity reduced the significance of movie_meter_IMDBpro (popularity ranking), this predictor was retained in the final model due to its positive impact on the adjusted R-squared, indicating an improvement in model performance.

c. Calculation of min_star_meter (most popular actor ranking)

   This predictor was derived by selecting the minimum value among log_actor1_star_meter, log_actor2_star_meter, and log_actor3_star_meter. The rationale behind this approach is that a lower star_meter ranking indicates greater popularity, as the star_meter is a 2022 IMDbPro ranking, where lower values correspond to more famous actors. Therefore, the minimum value among the three actors' rankings is used to represent the highest popularity in the movie's cast.

d. Polynomial Degree Selection

   Residual plots (see Appendix 5) indicated non-linearity in all numerical predictors except for log_movie_budget and log_duration. For the non-linear predictors, polynomial transformations

were applied, and the optimal polynomial degree (ranging from 1 to 10) was selected based on the model that minimized MSE.

Categorical Predictors

The selected categorical variables are release_month, country, maturity_rating, aspect_ratio, and dummy variables for all genres. Only those with a p-value below 0.05 in simple linear regression (see Appendix 6) were retained. However, due to the high correlation between country and language, as well as between release_year and colour_film, we opted to exclude language and colour_film from the final model to avoid multicollinearity. Additionally, to address the issue of category imbalance, categories with fewer than 50 observations were reclassified as "other," ensuring more reliable estimates.

Added Additional Dummy Variables

In addition to the provided predictors, we constructed and tested several new variables to capture specific industry-related dynamics that might affect movie performance. After evaluating their impact in the final regression model, three self-developed dummy variables were retained due to their positive contribution to the adjusted R-squared.

a. top_80_distributor: differentiates movies distributed by the top 80 distributors from others.
   This dummy variable aims to capture the added value of being associated with an influential distributor. Movies handled by larger distributors are more likely to receive extensive promotional campaigns, leading to better financial outcomes and stronger industry positioning, thus justifying the variable's inclusion.
b. top_100_director: distinguishes films directed by the top 100 directors.
   Director reputation often correlates with higher-quality production, a larger following, and greater trust from studios and audiences alike. Famous directors also tend to attract better actors, larger budgets, and more robust marketing strategies, which collectively enhance a movie's commercial and critical success. This variable helps capture the influence of a top-tier director on a movie's performance.
c. nb_faces_three: identifies films with more or equal to three faces.
   The number of featured faces often reflects the prominence of the movie's cast. Specifically, promotional materials with exactly three lead actors tend to be a strategic marketing choice, signaling a strong ensemble cast. This predictor allows us to quantify and assess the marketing impact of featuring three lead actors, capturing the potential influence of cast visibility on movie performance.

As a result, our final selected model incorporates a diverse range of processed predictors. These predictors were chosen based on their contribution to improving model performance and minimizing

error. For a detailed summary of the regression model and results, please refer to the stargazer table in Appendix 7.

4. Results

Final Model Results: Refer to Appendix 7: Stargazer Table for the Final Model for further details about the final model selected.

a. $\mathbf{R^2 = 0.5124,}$ adjusted and $\mathbf{R^2 = 0.4988.}$ In terms of predictability, approximately 51.24% of the variance in the IMDb score can be explained by the independent variables selected for the model and when penalizing the addition of the predictors with low relevance, it can be said that almost 50% of the variance is explained by the predictors.
b. Predictive Power: Our model has a Mean squared error **(MSE) of 0.6043417.**
c. Significance of predictors (F-stat P-value): P-value about 0, F-statistic for joint Hypothesis for all β is 37.79; speaks to the significance of the predictors in predicting ratings.
d. Statistically we can say that the following predictors are the most significant in the model selected: At **1% level of significance,** log_movie_budget, log_duration, countryUSA, maturity_ratingR, horror, drama, animation, crime, action, sport, poly(release_year,3)1, poly(nb_news_articles, 4)1, poly(nb_news_articles, 4)2, poly(nb_news_articles, 4)4, poly(log_movie_meter_IMDBPro, 4)1, poly(log_movie_meter_IMDBPro,4)2, poly(log_movie_meter_IMDBPro,4)3, poly(min_star_meter, 3)2, poly(min_star_meter, 3)3, top_100_director, are significant.
    i. At **5% level of significance**, aspect_ratio2.35, war, romance, poly(release_year,3)3, poly(nb_faces,2)2, nb_faces_three, are significant.
    ii. At 10% level of significance, maturity_ratingPG and western are significant.

Based on the results of the selected regression model, it appears that movies with higher budgets do not necessarily receive better IMDb scores. Specifically, the coefficient for the logarithm of the movie budget is -0.184, indicating that, holding other variables constant, a one-unit increase in log budget is associated with a 0.184-point decrease in IMDb score. In contrast, longer movies tend to have higher ratings: the coefficient for the logarithm of movie duration is 1.323, meaning that a one-unit increase in log duration corresponds to a 1.323-point increase in IMDb score.

Regarding the release month, movies released in November have a positive coefficient of 0.104, suggesting they receive higher IMDb scores compared to movies released in October (0.054) and December (0.033). This indicates that, all else being equal, movies released in November tend to score higher than those released in October and December.

The country of origin also affects IMDb scores. Films from the United States have a negative coefficient of -0.185, indicating they are associated with lower scores compared to other countries. In contrast, movies from the United Kingdom have a positive coefficient of 0.099, suggesting they tend to receive slightly higher scores.

Also, IMDb score is affected negatively if the movie belongs to musical, horror, action, adventure, romance or thriller genre. However, genres such as animation, drama, western, crime, western, war, sport or science fiction tend to impact positively in the IMDb score of the film. Refer to Appendix 7: Stargazer Table for the Final Model for further details about the coefficients.

The final selected model includes polynomial terms for the release year, with high and negative coefficients. This implies that, after accounting for other variables, more recent movies tend to have lower IMDb scores than older ones.

The number of articles about a movie impacts its IMDb score in a complex way, exhibiting both positive and negative effects, which indicates a non-linear relationship. This is also the case for variables such as IMDbPro metrics, the number of faces in the movie poster, and the minimum star meter rating.

Lastly, movies with a maturity rating of 'R' tend to have higher IMDb scores compared to other maturity rating levels. An aspect ratio of 2.35 has a positive coefficient of 0.081, indicating a slight increase in IMDb score associated with this aspect ratio. Additionally, movies featuring fewer than three faces on their poster have a positive coefficient of 0.184, suggesting they tend to receive higher ratings. Having a director who is listed among the top 100 directors also tends to increase the IMDb score (0.178).

Predictions (12 movies)

a. **Venom: The Last Dance (Predicted IMDb Score 5.80):** the prediction for this film is reasonable since the movie has a high budget, which it is a sign of lower the IMDb score. Also, it was produced in the United States, and it is an action/adventure film, which tends to decrease the score as well. It is going to be released in October, the maturity rate is PG-13, and the aspect ratio is 2.35 which are variables that do not have high impact on the IMDb score since their coefficient are lower than 0.1. However, it has a considerable value duration (109 mins) which might have balanced the score.
b. **Your Monster (Predicted IMDb Score: 5.):** it is going to be released in October, it was produced in the United Kingdom and the movie poster does not have any face in it. These last three factors contribute to increase the score slightly. However, there are other factors such as the release year, that fact that it belongs to the horror and romance genre, the low value of number of articles in the news (81) of the main country of the film, that might have affected negatively the predicted score, lowering down to the presented predicted value.
c. **Hitpig! (Predicted Score: 5.8):** it has a low duration (86mins), and it is also cataloged as an action and adventure film, so as mentioned before, this contributes to reducing the predicted score.

However, it is going to be released in November, it was also produced in the United Kingdom, the movie poster has less than three faces and it is an animated movie; these are factors that help to increase the IMDb score significantly.

d. **A Real Pain (Predicted IMDb score: 6.6):** this is the movie with the one of highest IMDd predicted score. It is also going to be released in November, it belongs to the drama genre, and it has less than three faces in its poster, so these factors contribute to increase the IMDb score by more than 0.5 points compared to previous movies. Nevertheless, the movie was produced in the United States, so this impacts negatively on the rating score.
e. **Elevation (Predicted IMDb score: 5.2):** this is the movie with the lowest predicted IMDb score! It was also made in the United States, it has a short duration (90 mins), it is an action and thriller movie, its director and distributors are not on the top 100 and top 80 list respectively. Hence, this film has a lot of factors that may contribute to not being a well-rated movie at all.
f. **The Best Christmas Pageant Ever (Predicted IMDb score: 5.7):** it is going to be released this November, does not have an extremely high budget and it is a drama movie, but the fact that has too many faces in the movie poster, and it was produced in the United States, affects its IMDb score negatively.
g. **Kanguva (Predicted IMDb score: 5.6):** Kanguva fares a low rating according to our model, as the key factors bringing its rating down would include the proxy for press hype, i.e., number of articles in the news, it is cataloged as an action movie, the fact that it is not distributed by one of the top distributors in this dataset, the popularity of its most famous actor being low by the star meter, and its popularity on the IMDB movie meter ranking. Kanguva does have some factors that could steer its success such as release month (as movies in November tend to do better than all other months), its duration being long at 2 hours and 26 minutes.
h. **Red One (Predicted IMDb score: 5.7):** Red One does come in with a star cast of Dwayne Johnson, Chris Evans, and Lucy Liu, and is a holiday movie (holiday movies fare higher scores, on average), releasing during November but certain factors bring down its rating such as its genres of action and adventure. With a super high budget, this could be a setup for failure as movies with higher budget do not translate to better quality movies, which suggests remaining true in this case. With a director out of the top 100 directors, this movie will not shake up the box office.
i. **Heretic (Predicted IMDb score: 5.6):** Heretic is A24's most recent horror/thriller film. A24 has seen a lot of success at the box office with movies like "We live in time". Although, A24's adventure into horror has not been as successful in the past with "Lamb" and it does not seem to change much with Heretic. Horror films fail to significantly captivate the audience, as evidenced by the results of our model, thus venturing into this genre, as a US based production company, does not yield a high IMDb audience score. Additionally, the fact that it does not have a significant number of articles in the news about it contributes to have a predicted low rating.

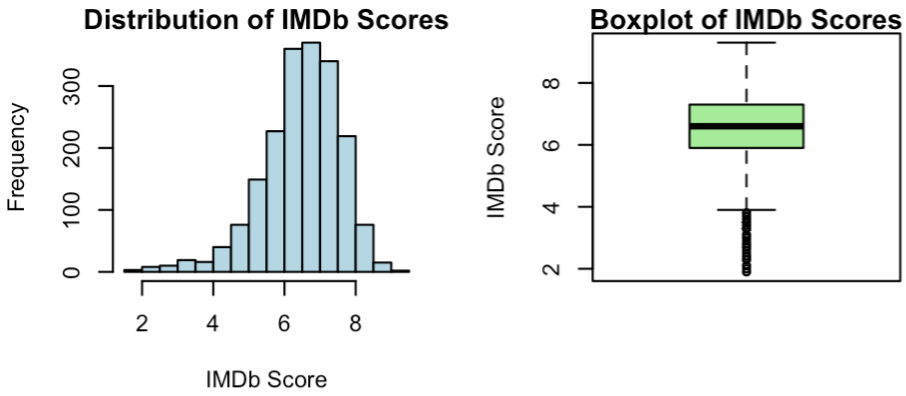
j. **Bonhoeffer Pastor. Spy. Assassin (Predicted IMDb score: 6.2):** For a war-era movie, Bonhoeffer's decision to keep it PG-13 might help in reaching a greater audience but would have benefited more from an R rating as the movie could have been more descriptive in displaying war themes. This can be seen in our model as PG-13 movies do not do as well as R rated movies. Bonhoeffer also suffers from a low presence in the press with very low articles in the news. Along with this, something that could also bring down the ratings could be not hiring famous actors for roles in this movie. However, it is catalog as a drama movie, which rises its score up.

k. **Gladiator II (Predicted IMDb score: 6.1):** The biggest upset of the year, behind Joker Folie a Deux, would be Gladiator II. With a star cast set to attract big numbers at the box office, the fact that it is a high budget sequel slightly lowers down its ratings. The number of articles about this movie are also low, being a movie from the United States, having 7 faces on its poster, and being an action/adventure movie also brings down its rating.

l. **Wicked 2 (Predicted IMDb score: 6.7):** Wicked 2 does the best in selected model for reasons where it outperforms the other movies in this list. It starts with a high budget, a long movie running beyond 2.5 hours, great marketing with about 1979 articles in the news mentioning the movie, a high popularity on IMDB pro movie meter, and star cast with Jeff Goldblum, Jonathan Bailey, and Cynthia Erivo. Factors that hurt this movie in the ratings would be the fact that it is a romantic movie, as movies in this genre tend to decrease its ratings in 0.093 compared to other genres, on average. Since this movie is based in the United States, that also holds back its rating in our model.

5. Appendices

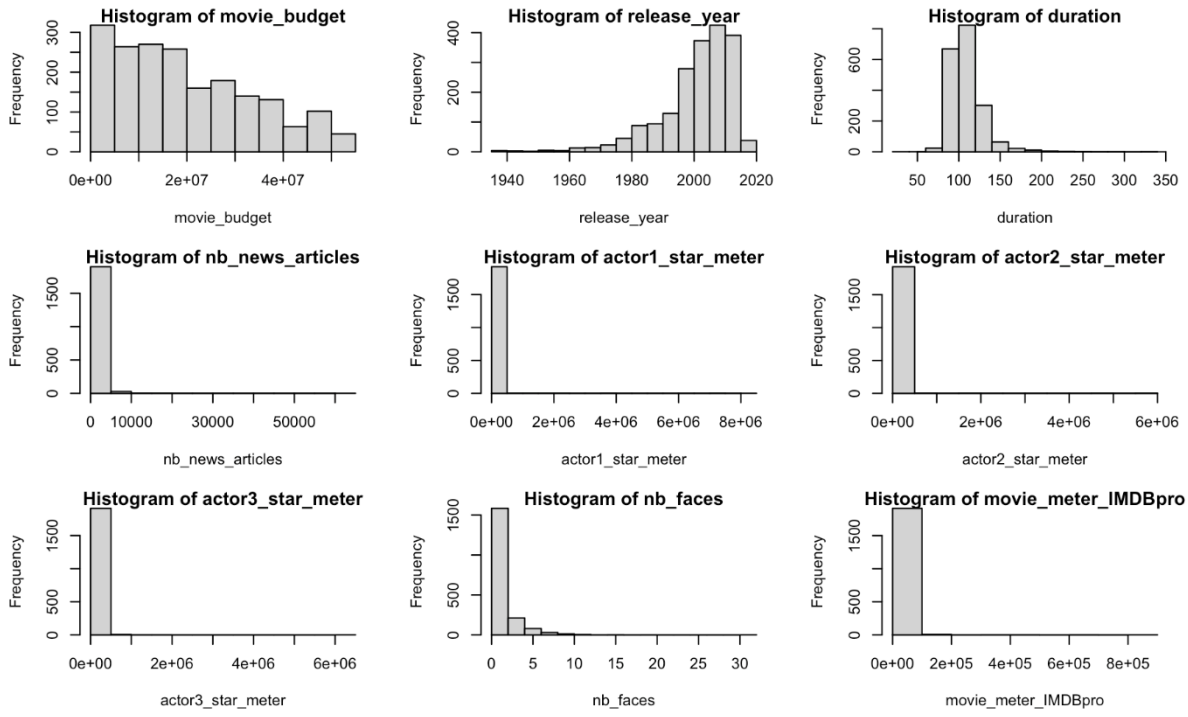
Appendix 1: IMDb scores Distributions

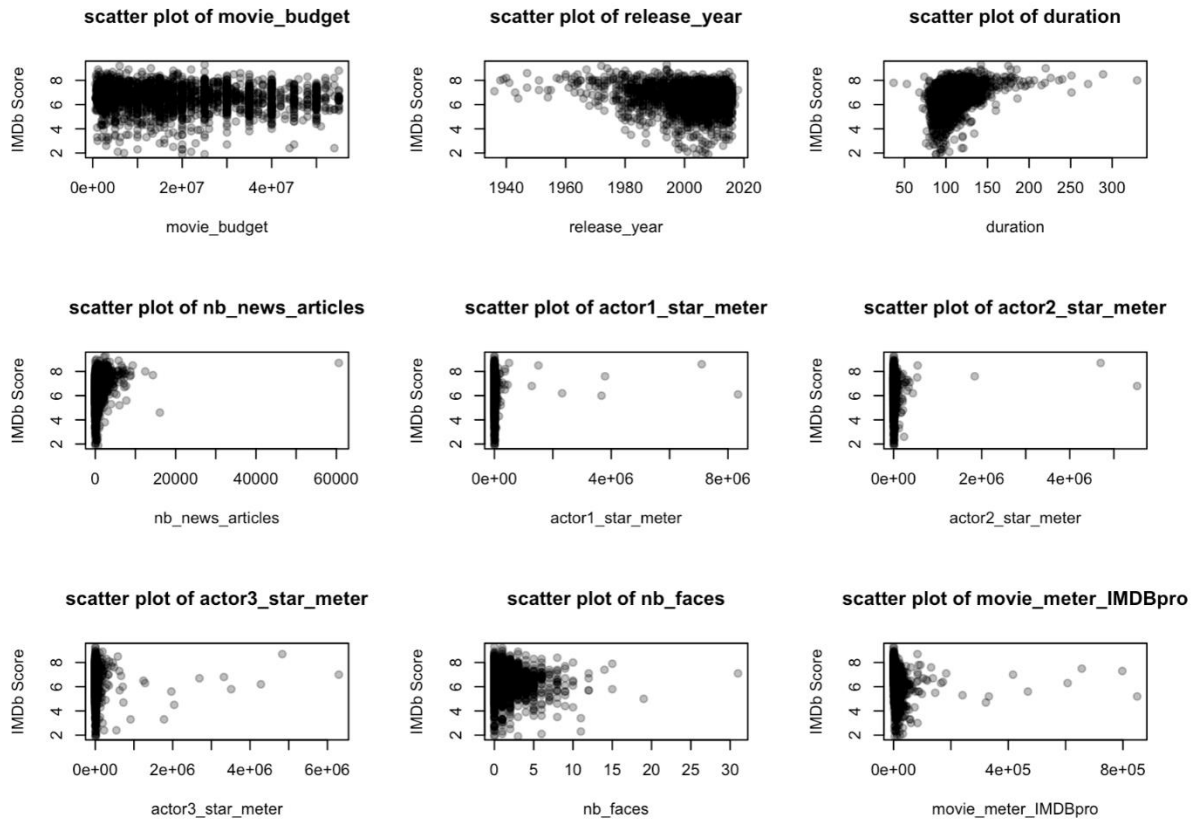


Appendix 2: Correlation Coefficients between imdb_score and Each Numeric Variable

```
[1] "Corr imdb_score, movie_budget : -0.08"
[1] "Corr imdb_score, release_year : -0.19"
[1] "Corr imdb_score, duration : 0.41"
[1] "Corr imdb_score, nb_news_articles : 0.23"
[1] "Corr imdb_score, actor1_star_meter : 0.03"
[1] "Corr imdb_score, actor2_star_meter : 0.04"
[1] "Corr imdb_score, actor3_star_meter : 0"
[1] "Corr imdb_score, nb_faces : -0.09"
[1] "Corr imdb_score, movie_meter_IMDBpro : -0.09"
```

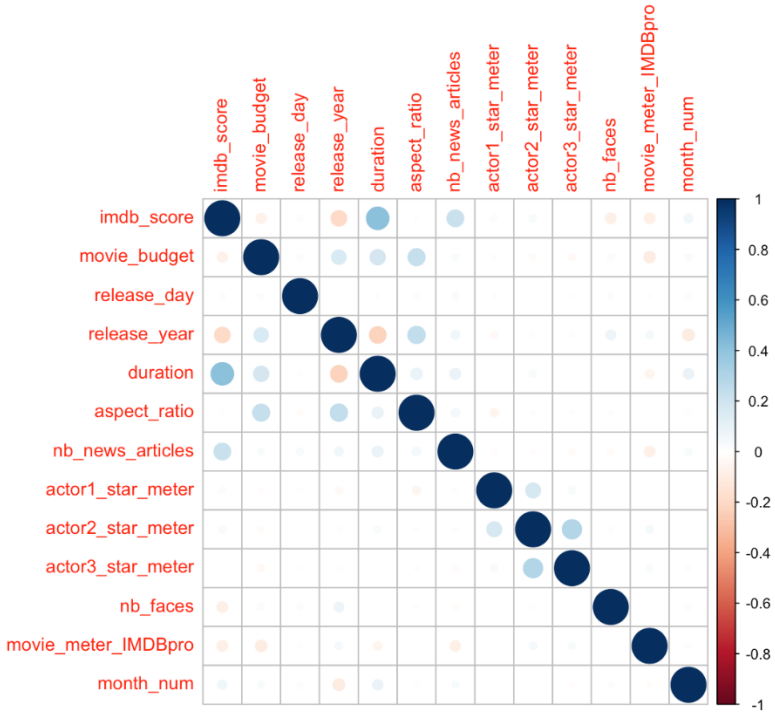
Appendix 3: Histograms and Scatter Plots for Numerical Predictors





Appendix 4: Correlations for Numerical Predictors

Correlation Matrix:



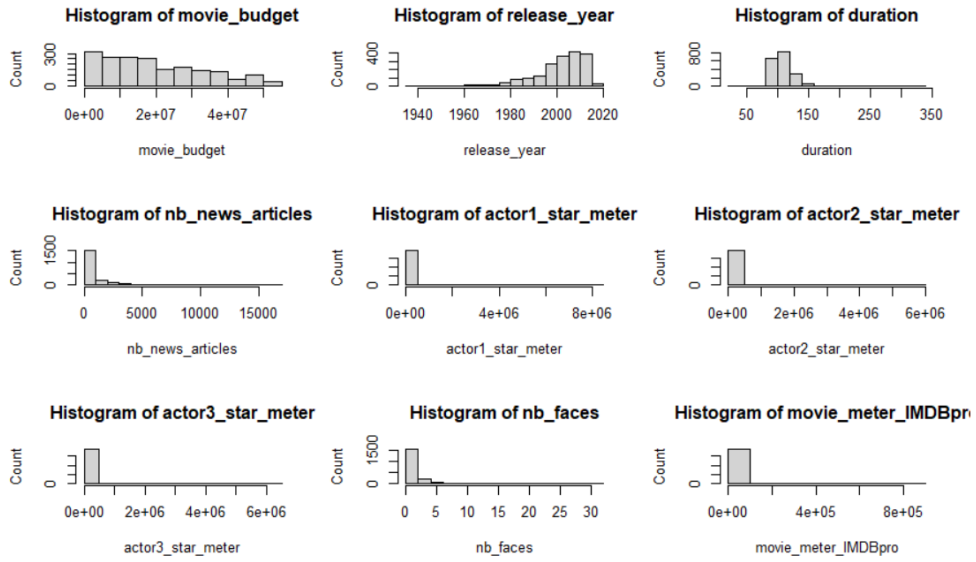
	movie_budget	release_year	duration	nb_news_articles	actor1_star_meter	actor2_star_meter	actor3_star_meter	nb_faces	movie_meter_IMDBpro
movie_budget	1.00	0.17	0.19	0.03	-0.02	-0.03	-0.03	0.03	-0.10
release_year	0.17	1.00	-0.22	0.06	-0.03	0.02	0.02	0.07	0.04
duration	0.19	-0.22	1.00	0.09	0.00	0.03	-0.01	0.01	-0.06
nb_news_articles	0.03	0.06	0.09	1.00	-0.02	-0.02	-0.03	-0.03	-0.09
actor1_star_meter	-0.02	-0.03	0.00	-0.02	1.00	0.18	0.04	0.00	0.01
actor2_star_meter	-0.03	0.02	0.03	-0.02	0.18	1.00	0.30	-0.01	0.04
actor3_star_meter	-0.03	0.02	-0.01	-0.03	0.04	0.30	1.00	0.00	0.03
nb_faces	0.03	0.07	0.01	-0.03	0.00	-0.01	0.00	1.00	0.00
movie_meter_IMDBpro	-0.10	0.04	-0.06	-0.09	0.01	0.04	0.03	0.00	1.00

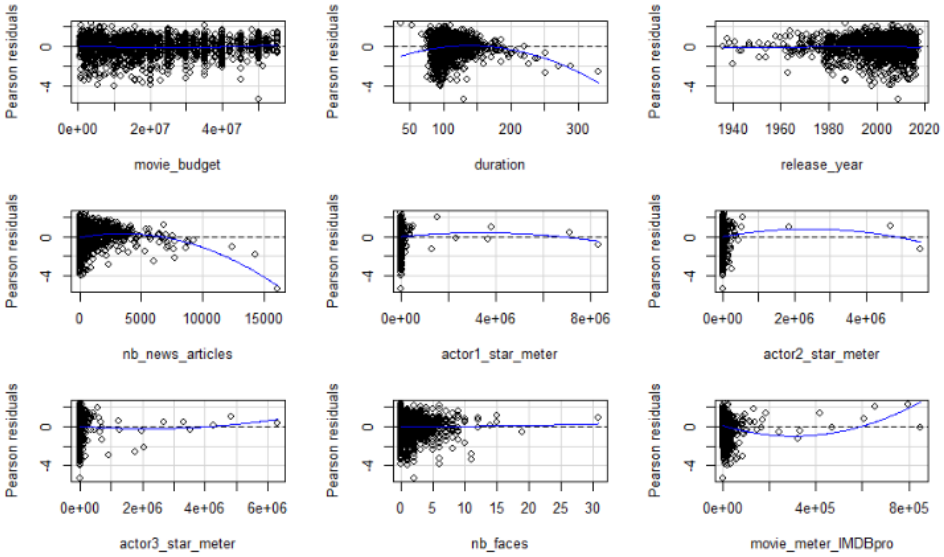
VIF test results:

```
       movie_budget        release_year            duration    nb_news_articles
           1.102028            1.125587            1.130042            1.025522
  actor1_star_meter   actor2_star_meter   actor3_star_meter            nb_faces
           1.035145            1.137935            1.100340            1.007940
movie_meter_IMDBpro
           1.023726
```

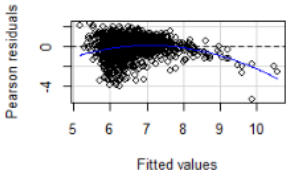
Appendix 5: Residual Plots and Histograms Before and After Log Transformation

Before Transformation:

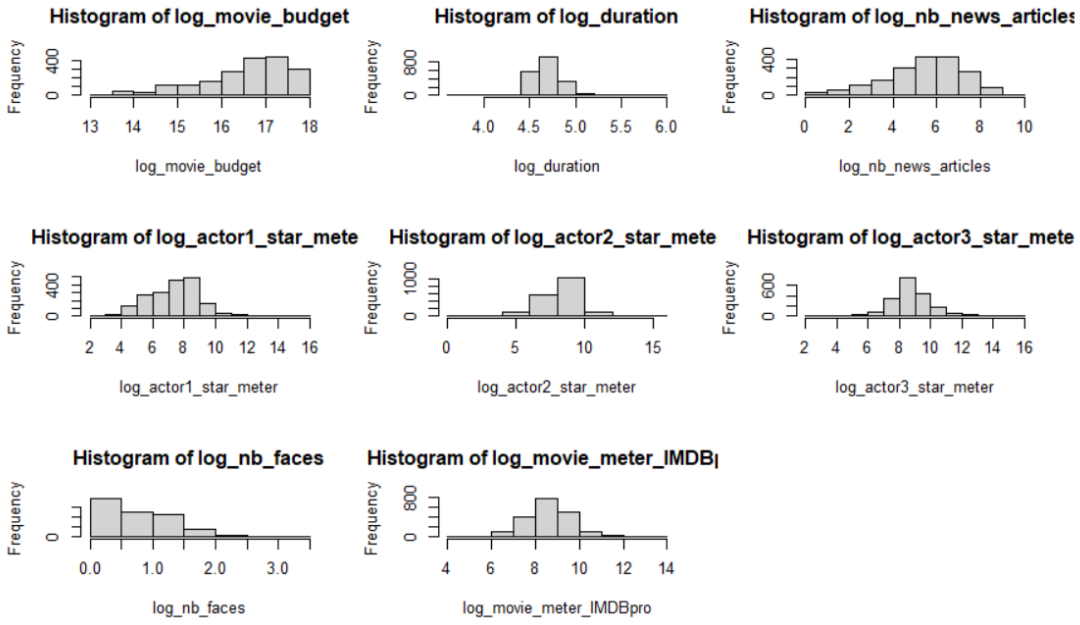




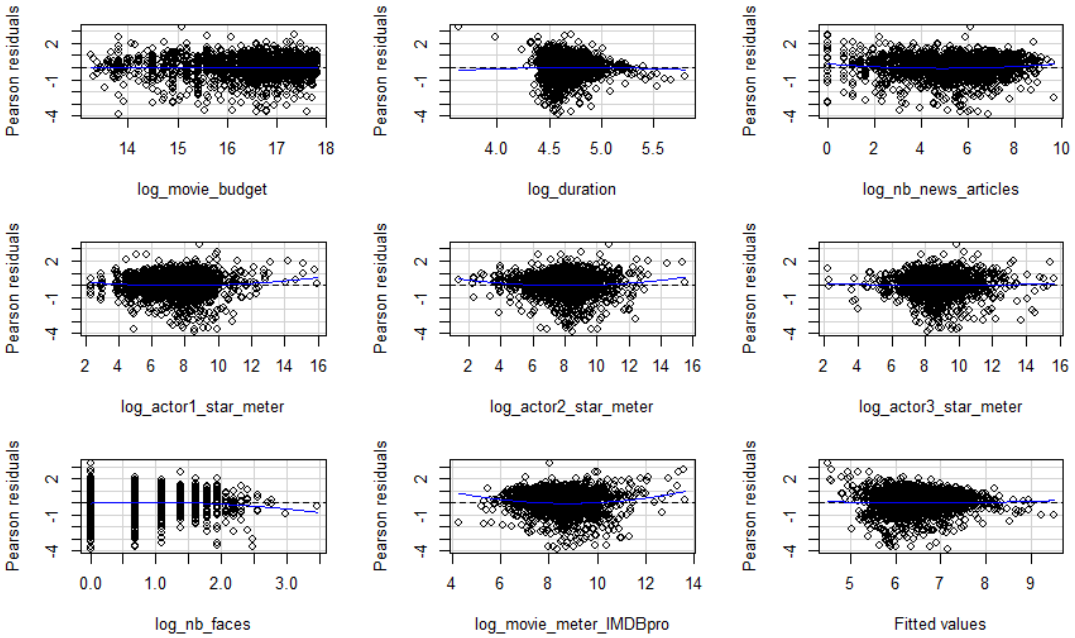

```
                  Test stat Pr(>|Test stat|)
movie_budget          2.9713         0.003003 **
duration             -7.3678        2.568e-13 ***
release_year         -0.4214         0.673514
nb_news_articles     -9.6252        < 2.2e-16 ***
actor1_star_meter    -1.1260         0.260319
actor2_star_meter    -1.4455         0.148471
actor3_star_meter     1.1965         0.231649
nb_faces              0.3245         0.745582
movie_meter_IMDBpro   6.0303        1.959e-09 ***
Tukey test          -10.5415        < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



After Transformation:



```
                          Test stat Pr(>|Test stat|)
log_movie_budget            -0.4362        0.6627760
log_duration                -0.5800        0.5619992
log_nb_news_articles         3.3480        0.0008297 ***
log_actor1_star_meter        2.5291        0.0115154 *
log_actor2_star_meter        2.9248        0.0034876 **
log_actor3_star_meter        0.7412        0.4586362
log_nb_faces                -3.0302        0.0024767 **
log_movie_meter_IMDBpro      4.6678        3.257e-06 ***
Tukey test                   0.8878        0.3746423
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Appendix 6: Results of Simple Linear Regression for Categorical Predictors

i. release_day

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.27532    0.12210  51.394  < 2e-16 ***
release_day2  0.28579    0.19018   1.503  0.13308
release_day3  0.28068    0.19460   1.442  0.14938
release_day4  0.35593    0.19704   1.806  0.07102 .
release_day5  0.27653    0.19018   1.454  0.14610
release_day6  0.38926    0.19704   1.976  0.04836 *
release_day7  0.23044    0.19232   1.198  0.23097
release_day8  0.02468    0.19580   0.126  0.89973
release_day9  0.15195    0.18916   0.803  0.42192
release_day10 0.37536    0.17503   2.145  0.03211 *
release_day11 0.09293    0.18202   0.511  0.60973
release_day12 0.20882    0.17003   1.228  0.21954
release_day13 -0.08844   0.18365  -0.482  0.63018
release_day14 0.36595    0.18202   2.010  0.04452 *
release_day15 0.10498    0.17973   0.584  0.55923
release_day16 0.31261    0.18629   1.678  0.09349 .
release_day17 0.12320    0.17830   0.691  0.48966
release_day18 0.32629    0.18283   1.785  0.07447 .
release_day19 0.51504    0.16953   3.038  0.00241 **
release_day20 0.47065    0.16764   2.807  0.00504 **
release_day21 0.23968    0.17105   1.401  0.16132
release_day22 0.21515    0.16904   1.273  0.20326
release_day23 0.25039    0.17694   1.415  0.15721
release_day24 0.38929    0.18047   2.157  0.03113 *
release_day25 0.38247    0.15950   2.398  0.01659 *
release_day26 0.18041    0.18365   0.982  0.32605
release_day27 0.04920    0.19123   0.257  0.79698
release_day28 0.15558    0.18916   0.822  0.41090
release_day29 0.32876    0.19580   1.679  0.09331 .
release_day30 0.26786    0.20248   1.323  0.18604
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.071 on 1893 degrees of freedom
Multiple R-squared:  0.01799,   Adjusted R-squared:  0.002944
F-statistic: 1.196 on 29 and 1893 DF,  p-value: 0.2178
```

ii. country

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.800      1.061   4.524 6.45e-06 ***
countryAustralia           1.517      1.084   1.400  0.16168
countryBelgium             2.300      1.500   1.533  0.12549
countryBrazil              3.200      1.500   2.133  0.03309 *
countryCanada              1.218      1.075   1.134  0.25714
countryChina               2.850      1.300   2.193  0.02842 *
countryColombia            2.700      1.500   1.799  0.07212 .
countryCzech Republic      1.600      1.500   1.066  0.28643
countryDenmark             0.900      1.500   0.600  0.54872
countryFrance              1.795      1.074   1.671  0.09489 .
countryGeorgia             0.800      1.500   0.533  0.59400
countryGermany             1.827      1.077   1.697  0.08993 .
countryGreece              1.900      1.500   1.266  0.20559
countryHong Kong           1.725      1.186   1.454  0.14608
countryHungary             1.000      1.500   0.666  0.50522
countryIndia               2.600      1.500   1.733  0.08331 .
countryIndonesia           2.800      1.500   1.866  0.06220 .
countryIreland             2.220      1.162   1.910  0.05629 .
countryItaly               2.425      1.125   2.155  0.03130 *
countryJapan               1.060      1.162   0.912  0.36190
countryKyrgyzstan          3.900      1.500   2.599  0.00942 **
countryMexico              3.300      1.500   2.199  0.02798 *
countryNetherlands         1.600      1.300   1.231  0.21838
countryNew Zealand         2.600      1.162   2.237  0.02541 *
countryOfficial site       1.500      1.500   1.000  0.31761
countryPeru                0.600      1.500   0.400  0.68931
countryRussia              2.500      1.500   1.666  0.09586 .
countrySouth Africa        1.880      1.162   1.617  0.10594
countrySouth Korea         1.200      1.300   0.923  0.35590
countrySpain               2.114      1.134   1.864  0.06248 .
countryTaiwan              3.100      1.500   2.066  0.03897 *
countryUK                  2.158      1.064   2.028  0.04267 *
countryUSA                 1.673      1.061   1.577  0.11503
countryWest Germany        2.600      1.500   1.733  0.08331 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.061 on 1889 degrees of freedom
Multiple R-squared:  0.03901,   Adjusted R-squared:  0.02223
F-statistic: 2.324 on 33 and 1889 DF,  p-value: 3.253e-05
```

iii. release_month

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.41124    0.08178  78.398  < 2e-16 ***
release_monthAug -0.04970    0.11565  -0.430  0.66741
release_monthDec  0.48952    0.12375   3.956 7.91e-05 ***
release_monthFeb  0.11069    0.11823   0.936  0.34928
release_monthJan  0.16876    0.11046   1.528  0.12673
release_monthJul  0.07119    0.11968   0.595  0.55204
release_monthJun  0.24726    0.11990   2.062  0.03932 *
release_monthMar -0.06092    0.11784  -0.517  0.60521
release_monthMay -0.17755    0.13774  -1.289  0.19756
release_monthNov  0.32592    0.11968   2.723  0.00652 **
release_monthOct  0.20635    0.10918   1.890  0.05891 .
release_monthSep -0.01984    0.11298  -0.176  0.86059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.063 on 1911 degrees of freedom
Multiple R-squared:  0.02401,   Adjusted R-squared:  0.0184
F-statistic: 4.274 on 11 and 1911 DF,  p-value: 2.576e-06
```

iv. maturity_rating

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.5619     0.2293  32.974  < 2e-16 ***
maturity_ratingG      -0.9913     0.2917  -3.399 0.000691 ***
maturity_ratingGP     -0.2119     0.7777  -0.272 0.785285
maturity_ratingM      -0.1119     0.7777  -0.144 0.885601
maturity_ratingNC-17  -1.5286     0.6486  -2.357 0.018546 *
maturity_ratingPassed -0.3619     0.5733  -0.631 0.527962
maturity_ratingPG     -1.0955     0.2387  -4.590 4.72e-06 ***
maturity_ratingPG-13  -1.2988     0.2335  -5.563 3.02e-08 ***
maturity_ratingR      -0.8952     0.2317  -3.864 0.000115 ***
maturity_ratingTV-14  -2.6619     0.6486  -4.104 4.24e-05 ***
maturity_ratingTV-G   -2.2952     0.6486  -3.539 0.000412 ***
maturity_ratingX      -0.7619     0.4366  -1.745 0.081154 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.051 on 1911 degrees of freedom
Multiple R-squared:  0.04626,   Adjusted R-squared:  0.04077
F-statistic: 8.426 on 11 and 1911 DF,  p-value: 1.205e-14
```

v. aspecto_ratio

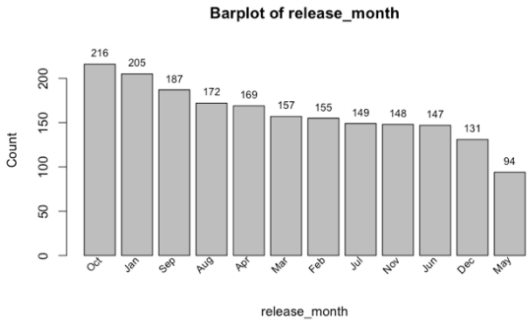
```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.500      1.064   7.988 2.35e-15 ***
aspect_ratio1.33   -2.227      1.111  -2.004  0.04521 *
aspect_ratio1.37   -1.450      1.083  -1.339  0.18075
aspect_ratio1.5    -1.300      1.505  -0.864  0.38777
aspect_ratio1.66   -1.365      1.095  -1.246  0.21278
aspect_ratio1.75   -0.800      1.303  -0.614  0.53939
aspect_ratio1.78   -2.906      1.093  -2.658  0.00793 **
aspect_ratio1.85   -2.027      1.065  -1.904  0.05704 .
aspect_ratio2.2    -0.800      1.149  -0.696  0.48649
aspect_ratio2.35   -1.953      1.065  -1.834  0.06674 .
aspect_ratio2.39   -1.814      1.138  -1.595  0.11091
aspect_ratio2.4    -1.267      1.229  -1.031  0.30273
aspect_ratio2.55   -1.900      1.505  -1.263  0.20690
aspect_ratio2.76   -0.900      1.505  -0.598  0.54987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.064 on 1909 degrees of freedom
Multiple R-squared:  0.02322,   Adjusted R-squared:  0.01657
F-statistic: 3.491 on 13 and 1909 DF,  p-value: 2.108e-05
```

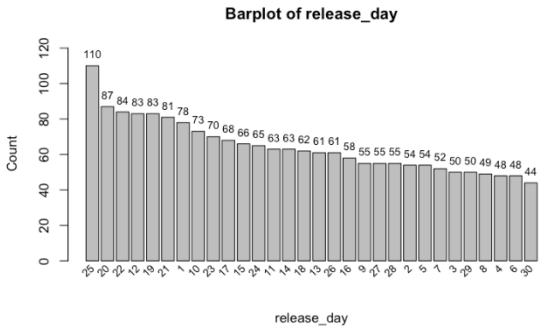
Appendix 7: Stargazer Table for the Final Model

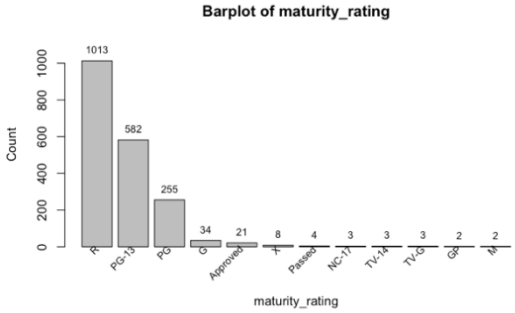
	Dependent variable:
	IMDB Score
log_movie_budget	-0.184***
	(0.023)
log_duration	1.323***
	(0.132)
release_monthAug	-0.014
	(0.084)
release_monthDec	0.033
	(0.091)
release_monthFeb	0.031
	(0.086)
release_monthJan	0.039
	(0.080)
release_monthJul	-0.010
	(0.087)
release_monthJun	0.018
	(0.088)
release_monthMar	-0.051
	(0.085)
release_monthMay	-0.103
	(0.099)
release_monthNov	0.104
	(0.087)
release_monthOct	0.054
	(0.079)
release_monthSep	-0.063
	(0.082)
countryUK	0.099
	(0.081)
countryUSA	-0.185***
	(0.060)
maturity_ratingPG	0.198*
	(0.113)
maturity_ratingPG-13	0.137
	(0.114)
maturity_ratingR	0.292***
	(0.113)
aspect_ratio2.35	0.081**
	(0.040)
aspect_ratioothers	-0.011
	(0.091)
action	-0.310***
	(0.053)
adventure	-0.077
	(0.062)
scifi	0.050
	(0.063)
thriller	-0.070
	(0.047)
musical	-0.117
	(0.072)
romance	-0.094**
	(0.045)
western	0.249*
	(0.136)
sport	0.262***
	(0.086)

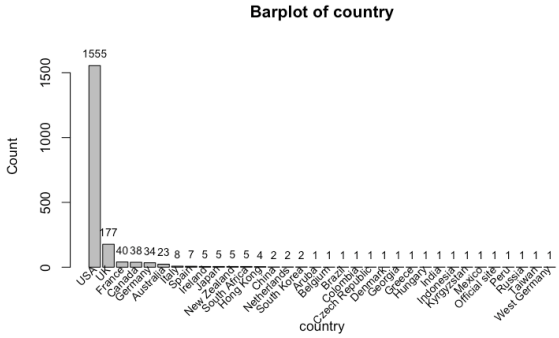
horror	-0.492***
	(0.065)
drama	0.396***
	(0.043)
war	0.194**
	(0.099)
animation	1.012***
	(0.184)
crime	0.145***
	(0.051)
poly(release_year, 3)1	-8.222***
	(1.057)
poly(release_year, 3)2	-0.594
	(0.933)
poly(release_year, 3)3	-2.061**
	(0.834)
poly(nb_news_articles, 4)1	5.824***
	(0.971)
poly(nb_news_articles, 4)2	-5.001***
	(0.851)
poly(nb_news_articles, 4)3	0.475
	(0.824)
poly(nb_news_articles, 4)4	-2.118***
	(0.795)
poly(nb_faces, 2)1	-1.474
	(1.259)
poly(nb_faces, 2)2	-1.843**
	(0.838)
poly(log_movie_meter_IMDBpro, 4)1	-14.219***
	(1.096)
poly(log_movie_meter_IMDBpro, 4)2	3.546***
	(0.828)
poly(log_movie_meter_IMDBpro, 4)3	3.724***
	(0.795)
poly(log_movie_meter_IMDBpro, 4)4	-0.942
	(0.786)
poly(min_star_meter, 3)1	0.883
	(0.868)
poly(min_star_meter, 3)2	2.990***
	(0.800)
poly(min_star_meter, 3)3	2.745***
	(0.782)
top_80_distributor	0.072
	(0.053)
top_100_director	0.178***
	(0.041)
nb_faces_three	0.184**
	(0.092)
Constant	2.840***
	(0.623)
Observations	1,923
R^2	0.512
Adjusted R^2	0.499
Residual Std. Error	0.760 (df = 1870)
F Statistic	37.785*** (df = 52; 1870)
Note:	*p<0.1; **p<0.05; ***p<0.01

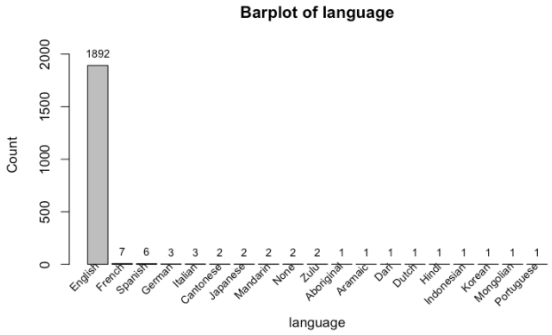
Appendix 8: Bar Charts for Categorical Variables

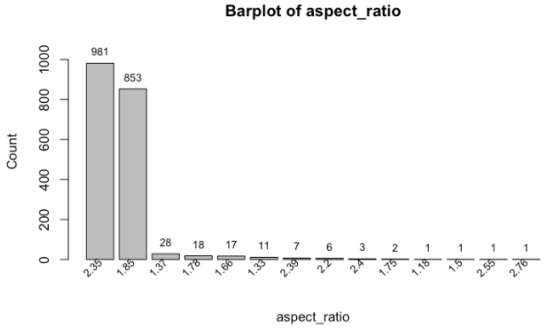


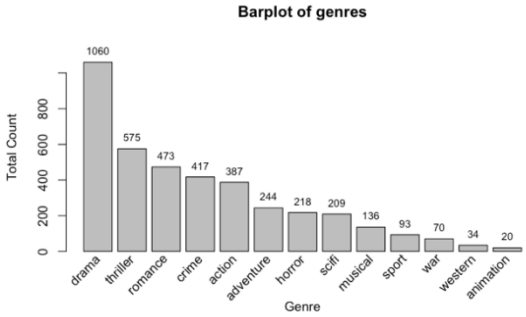


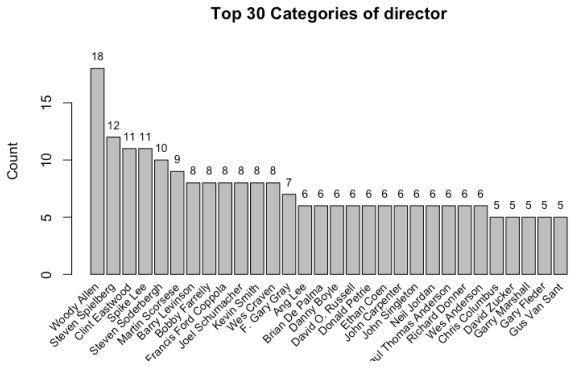


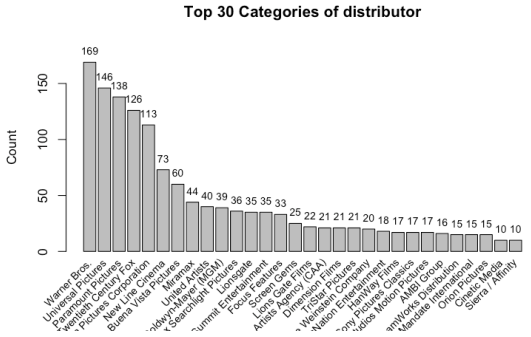


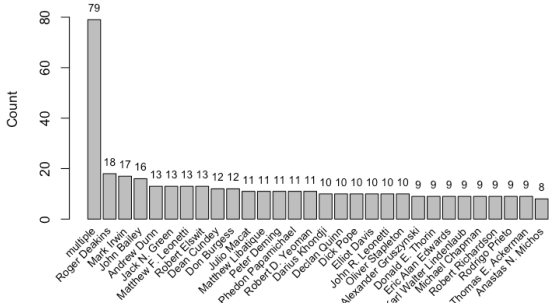


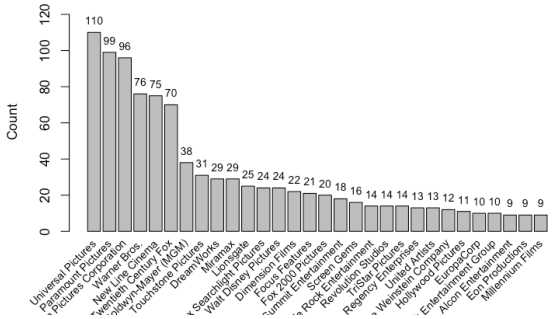


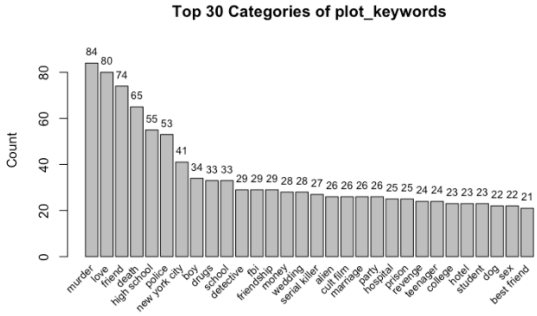












6. Code

Please refer to midterm_Wicked 6_Rcode.Rmd.