1 Introduction

In the fascinating world of data science and predictive analytics, few challenges light the fire of curiosity quite like the quest to foresee the destiny of upcoming blockbuster films. In this project, we employ the tools of statistical modeling, a sprinkle of creativity, and the invaluable lessons of trial and error to forecast the IMDb ratings of twelve eagerly anticipated movies destined for the silver screen in the near future. This journey allows us to put into practice the knowledge and skills we have gathered in our courses and delve into the practical world of predictive data analytics.

The central objective of our research is to forecast the IMDb ratings of upcoming blockbuster films. IMDb ratings are renowned for their natural variability and susceptibility to change over time, establishing them as a reliable metric for assessing the audience's perception of a film's q uality. While these ratings may experience initial fluctuations in the days following a movie's release, they inevitably settle, forming a sturdy foundation for our predictive model.

The dataset includes a wide range of information, such as movie titles, IMDb ratings, budgets, release details, language, production country, and more. Nonetheless, not all variables contribute significantly to the final prediction. Our secondary objective is to refine the dataset by filtering or modifying it, retaining the most meaningful and influential predictors. This process ensures the creation of a model that can yield the most accurate results.

In our pursuit of precise predictions, we address not only model fitness but also the challenges of data relationships, such as non-linearity, heteroskedasticity, the influence of outliers, and the presence of c ollinearity. We employ various statistical techniques, such as feature selection and cross-validation, to mitigate these issues. Our approach involves continuous experimentation to fine-tune the model.

In our quest to forecast IMDb ratings for forthcoming blockbusters, we navigate the dynamic world of data analytics, armed with statistical modeling, creativity, and determination. By refining our dataset and addressing challenges, we're poised to unlock the secrets of cinematic success. This journey embodies our dedication to harnessing the power of data to predict the unpredictable.

2 Data Exploration and Data Pre-processing

We explored the data by checking the distributions of the categorical and numerical variables with histograms, boxplots, *table()*, and *skewness()* functions. Please find the histogram and boxplot diagrams in the Appendix - Variable Distribution Plots.

Then we preprocessed the data by dropping excessive labels. We found that there are three labels in the dataset, so we dropped *movie_title* and *imdb_link* for simplicity. We also handled the skewness by re-categorization and log transformation. Notable findings and necessary data preprocessing processes are listed in Section 2.1 and Section 2.2.

2.1 Categorical Variables

We examined and preprocessed the distribution of categorical variables:

1. *release_month:* Each month has an approximately similar number of movie releases, with slightly more movie releases in January and October.

2. *language:* There are 19 distinct languages in this dataset. English is the dominant language which takes 98% among all observations. To address the skewness, we merged non-English movies into a category called "other languages".

3. *country:* There are 34 countries in the dataset. USA and UK together take over 89% among all observations. To address the skewness, we merged non USA and UK countries into a category called "other countries".

4. *maturity_rating:* There are 12 maturity ratings in the dataset: Approved, G, GP, M, NC-17, Passed, PG, PG-13, R, TV-14, TV-G, and X. Most of the films fall into PG, PG-13, and R. To address the skewness, we merged other infrequent ratings into a category called "other ratings".

5. ***Production and Cast Characteristics**:* This group includes *distributor*, *director*, *actor1*, *actor2*, *actor3*, *cinematographer*, and *production_company*. Each of these 7 categorical variables has more than 300 distinct categories. If we include these variables directly into our model, the overfitting problem might occur. As a result, we kept Warner Bros, Universal Pictures, Paramount Pictures, Twentieth Century Fox, and Columbia Pictures Corporation for distributor and merged other distributors into a new category called "other distributors". Similarly, we kept Warner Bros, Universal Pictures, Paramount Pictures, Twentieth Century Fox, Columbia Pictures Corporation, and New Line Cinema for *production_company* and merged other production companies into a new category called "other production companies". We also re-categorized cinematographer and director into binary variables, with 1 indicating the top 25% most frequently appearing names and 0 indicating the remaining 75%.

6. *colour_film:* There are two categories in this variable: color and black-and-white. On average, color films scored 0.76 points lower than black and white films.

7. ***Movie Genres**:* This group includes a list of binary variables: *action*, *adventure*, *scifi*, *thriller*, *musical*, *romance*, *western*, *sport*, *horror*, *drama*, *war*, *animation*, and *crime*. The genre of the movie explains a certain degree of variation in *imdb_scores*. For instance, we found that on average, action films scored 0.38 points lower than non-action films, and drama films scored 0.49 points higher than non-drama films.

2.2 Numerical Variables

We then explored and pre-processed the distribution of numerical variables.

1. *imdb_score:* The data varies from 1.9 to 9.5, with a mean of 6.51. It appears to be left-skewed with a concentration of movies having higher IMDb scores.

2. *movie_budget:* The data is ranging from $0.56 million to $300 million, has a mean of approximately $20.97 million, and right-skewed, with a few movies having significantly higher budgets than the rest.

3. *release_day:* It is uniformly distributed across the days of the month, with slight variations.

4. *release_year:* It is left-skewed, indicating an increasing number of movies being released in recent years, ranging from 1936 to 2017, with a mean release year around 2001.

5. *duration:* The data varies from 37 to 330 minutes, with an average duration of approximately 109.7 minutes and is right-skewed with a few movies having unusually long durations.

6. *aspect_ratio:* Majority of movies have an aspect ratio around 2.35 and 1.85, with a few exceptions, indicating a clustering of values around the mean. Thus, we treated this variable as categorical data and factorized it.

7. *nb_news_articles:* Right-skewed, with most movies having a low number of news articles, and a few having a very high number. The number of news articles related to the movies shows an extreme range from 0 to 1,739,000, with a mean of 770.6.

8. ***Actor Star Meter**:* This group includes *actor1_star_meter*, *actor2_star_meter*, *actor3_star_meter*. These variables are highly right-skewed, indicating that most actors have smaller star meter values (more famous), while a few have very high (less popular) values.

9. *nb_faces:* Right-skewed, with most movies having a low number of faces on their posters, ranging from 0 to 31, with a mean of 1.44.

10. *movie_meter_IMDBpro:* The data ranges from 71 to 8,495,500, with a mean of 11,612 and an extremely right skewness, suggesting that most movies are not highly ranked.

Besides visualization with histograms and boxplots, we also examined the skewness of numerical data using the skewness() function. Based on the skewness value, each variable is categorized as "Symmetric," "Moderately Skewed," or "Highly Skewed." Variables such as *actor1_star_meter*, *actor2_star_meter*, *actor3_star_meter*, *nb_news_articles*, and *movie_meter_IMDBpro* exhibit the highest skewness and therefore require careful consideration when used as predictors in our model. Highly skewed features can potentially bias the model's performance, leading to inaccurate predictions. Please refer to the skewness table in the Appendix - Skewness. Therefore, we conducted a log transformation to reduce the impact of these highly skewed variables. Moreover, since *nb_news_articles* has 0 values, we log-transformed it as *nb_news_articles* + 1 to avoid getting negative infinity. In the Appendix - Log-transformation, you will find a comparison between two sample sets of histograms: one set represents the original data, and the other set represents the data after a log-transformation has been applied.

3 Model Selection

Based on exploratory data analysis, we first included all the predictors we found possibly significant into a multiple linear regression model and then used stepwise Akaike Information Criterion (AIC) method to identify the model that best balances goodness of fit with simplicity. This provide us with a foundation model to further fine-tune:

$$\begin{aligned}
\text{imdb\_score} = & b_0 + b_1 \cdot \text{movie\_budget} + b_2 \cdot \text{release\_year} \\
& + b_3 \cdot \text{duration} + b_4 \cdot \text{language} + b_5 \cdot \text{country} \\
& + b_6 \cdot \text{maturity\_rating} + b_7 \cdot \text{nb\_news\_articles} + b_8 \cdot \text{director\_top25p} \\
& + b_9 \cdot \text{actor1\_star\_meter} + b_{10} \cdot \text{colour\_film} + b_{11} \cdot \text{nb\_faces} \\
& + b_{12} \cdot \text{action} + b_{13} \cdot \text{musical} + b_{14} \cdot \text{romance} \\
& + b_{15} \cdot \text{western} + b_{16} \cdot \text{sport} + b_{17} \cdot \text{horror} \\
& + b_{18} \cdot \text{drama} + b_{19} \cdot \text{animation} + b_{20} \cdot \text{crime} \\
& + b_{21} \cdot \text{movie\_meter\_IMDBpro} + b_{22} \cdot \text{cine\_top25p}
\end{aligned} \tag{1}$$

We conducted a linearity test on the foundation model and each predictor, and found the model to be non-linear. Specifically, as shown in the Appendix - Linearity Visual Test, *duration*, *nb_news_articles*, *actor1_star_meter*, and *movie_meter_IMDBpro* show non-linearity. To address the non-linearity issue, we used the looping method to find the best polynomial combination of these four categories. We assigned a degree of 2 for duration, a degree of 2 for *nb_news_articles*, a degree of 5 for *actor1_star_meter*, and a degree of 4 for *movie_meter_IMDBpro*.

We then tested the model on Heteroskedasticity. Both visual funnel test and ncvTest confirmed the existence of heteroskedasticity. After correcting heteroskedasticity errors, we found that western variable was no longer significant. We ran a 10-fold Test on MSE with and without western category and decided to remove *western* from our model. We also checked the outliers of the dataset by running a Bonferroni Test on the model and removed the eight observations that were identified as outliers. We used VIF to examine Collinearity and as shown in Appendix - Collinearity, no significant collinearity was found.

After all the tests, we finalized our model as the following:

$$\begin{aligned}\text{imdb\_score} =& b_0 + b_1 \cdot \text{movie\_budget} + b_2 \cdot \text{release\_year}\\ &+ b_3 \cdot \text{language} + b_4 \cdot \text{country} + b_5 \cdot \text{maturity\_rating}\\ &+ b_6 \cdot \text{director\_top25p} + b_7 \cdot \text{colour\_film} + b_8 \cdot \text{nb\_faces}\\ &+ b_9 \cdot \text{action} + b_{10} \cdot \text{musical} + b_{11} \cdot \text{romance}\\ &+ b_{12} \cdot \text{sport} + b_{13} \cdot \text{horror} + b_{14} \cdot \text{drama}\\ &+ b_{15} \cdot \text{animation} + b_{16} \cdot \text{crime} + b_{17} \cdot \text{cine\_top25p}\\ &+ b_{18} \cdot \text{actor1\_star\_meter}^5 + b_{19} \cdot \text{duration}^2\\ &+ b_{20} \cdot \text{nb\_news\_articles}^2 + b_{21} \cdot \text{movie\_meter\_IMDBpro}^4\end{aligned} \tag{2}$$

4 Results

4.1 Model Statistics

4.1.1 Model Summary

In this study, we have implemented a model with a Linear Model (LM). The linear regression model explained approximately 50.4% of the variance in IMDb scores, as indicated by the Multiple R-squared value of 0.504. The Adjusted R-squared, which adjusts for the number of predictors in the model, was slightly lower at 0.4953, suggesting that our model is reasonably well-fitted to the data, though there is still room for improvement. The F-statistic of 58.38 with a p-value less than 2.2e-16 indicates that our model is statistically significant. Please refer to Appendix - Model Statistics for details.

In addition to the analysis on the summary statistics, an out-of-sample performance evaluation was also carried out to measure the predictive power of the model. It provides insights into how well the model is expected to perform on unseen data, ensuring its reliability and effectiveness in practical applications. The mean squared error of our model is 0.62, which means that on average, the squared difference between the actual and predicted IMDb scores is approximately 0.62.

4.1.2 Predictor Significance

The significance of each predictor was evaluated based on the t-values and corresponding p-values. The following are predictors showed significant relationships with the IMDb score:

1. **Movie Budget:** A negative coefficient suggests that higher budgets are associated with lower IMDb scores.
2. **Release Year:** Older movies tend to have higher IMDb scores in our model.
3. **Language (Other Languages):** Movies in languages other than English are predicted to have higher IMDb scores.
4. **Colour Film:** Black and white films are predicted to have higher scores.
5. **Number of Faces in Poster:** A negative coefficient indicates that more faces are associated with lower IMDb scores.
6. **Genre:** Action, Horror, and Romance genres are associated with lower scores, while Drama, Animation, Sport, and Crime are associated with higher scores.
7. **Actor 1's Star Meter:** The 2022 ranking of the main actor made by IMDbPro has a complex relationship with IMDb scores, with 5th degree polynomial terms being the most significant.
8. **Duration:** Longer movies are generally predicted to have higher IMDb scores.
9. **Number of News Articles:** Movies with more news articles are associated with higher IMDb scores.
10. **IMDbPro MovieMeter:** A higher IMDbPro movie meter score is associated with lower IMDb scores.
11. **Maturity Rating (R):** R-rated films are predicted to have higher IMDb scores.
12. **Director:** Films directed by the top 25% of directors who have the highest frequency of appearances in the training set are predicted to have higher scores.
13. **Cinematographer:** Films shot by the top 25% of cinematographers who have the highest frequency of appearances in the training set are predicted to have higher scores.
14. **Country (UK and USA):** UK films are associated with higher scores, while US films are associated with lower scores.

Please refer to Appendix - Model Statistics for more details.

4.2 Prediction Results

Using the finalized model, we generated the predicted IMDb score for all 12 blockbuster movies, as shown in Appendix - Model Prediction Result. Among the 12 movies, *Napoleon* receives the highest predicted IMDb score of 7.09, and *Pencils vs Pixels* receives the lowest predicted IMDb score of 5.39. The average predicted IMDb score is 6.39. The main reason for *Napoleon*'s high predicted score might be due to its duration, country, maturity rating, number of news articles, and IMDbPro movie meter. *Napoleon* has the longest duration (158 minutes) among the 12 movies and duration has a positive correlation with the IMDb score. *Napoleon* is also the only movie that was produced in the United Kingdom. All other 11 films were produced in the United States. According to our model, UK films usually score higher than USA films. *Napoleon* also has an R maturity rating, appears in a large number of news articles, and has a relatively low IMDbPro movie meter, which all correlate with higher IMDb scores. On the other hand, *Pencil vs Pixels* receives a low predicted score because of its duration, maturity rating, number of news articles, and IMDbPro movie meter. It has the shortest duration (72 minutes) among the 12 movies, a G maturity rating, the lowest number of news articles, and the highest IMDbPro movie meter, all leading to a low IMDb score.

4.3 Recommendations

According to our model, about 80% of predictors are significant, indicating a strong fit and reliable predictive power. For film industry stakeholders, we suggest focusing on these statistically significant predictors, which could influence their movies' IMDb scores significantly. These pivotal factors encompass movie budget, release year, movie duration, number of faces in the film's main poster, IMDbPro's ranking of movies, and movie genres, particularly action, romance, horror, drama, and animation.

Delving into the details, it is crucial for major stakeholders in the film industry to deeply consider the budget allocation for a movie. The data reveals a strong statistical relationship between the IMDb rating and the movie's budget, albeit the relationship is negative and the coefficient relatively minimal. Our interpretation is that films with limited budgets resonate closely with ”arthouse cinema”, usually emphasizing higher artistic standards. Conversely, films with gargantuan budgets may veer towards commercialization, potentially compromising their intrinsic artistic merit. Moreover, older movies are perceived more favorably, indicating a nostalgic value. We recommend filmmakers should consider remakes, sequels, or inspiration from older classics while being careful to preserve the essence that made them popular. According to the model, posters with fewer faces are associated with higher IMDb scores. Perhaps, minimalist designs or avoiding overcrowded visuals can make a poster more appealing or intriguing. The IMDbPro MovieMeter score reveals a complex relationship between a movie's popularity and its perceived quality. A lower IMDb Pro ranking, indicating popularity, doesn't guarantee a high IMDb score. Stakeholders must find a balance between mass appeal and film quality. While the IMDbPro MovieMeter score gauges popularity, excessive promotion can raise expectations, potentially leading to lower ratings if unmet.

Furthermore, the genre of a film plays a vital role in shaping its IMDb rating. Our empirical findings suggest that certain genres, namely action, romance, and horror, might adversely influence IMDb ratings. In contrast, genres like drama and animation appear to bolster a film's rating when other factors are held constant. Consequently, we strongly advocate for producers to judiciously select a film's genre or perhaps contemplate a blend of genres, ensuring the optimal reception and success of the movie.

4.4 Conclusion

In order to predict IMDb ratings of the 12 upcoming blockbuster movies, we meticulously undertook the following steps:

- Data Pre-processing
  - Drop Irrelevant Labels

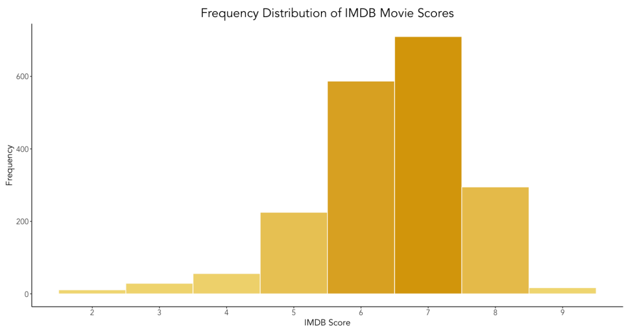
  - ○ Handling Categorical Data: Factorization and Skewness Management
  - ○ Numerical Data Processing: Addressing Skewness

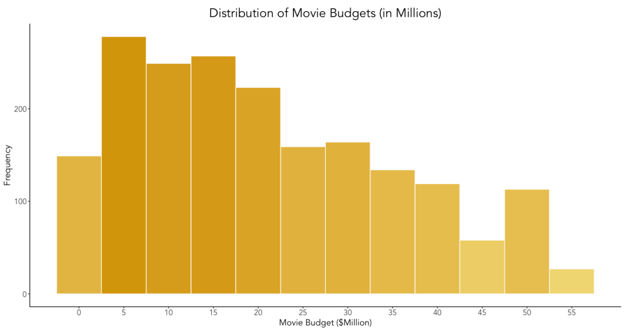
- Feature Selection and Model Building
  - ○ Linear Regression
    - Stepwise Variable Selection for Linear Regression
    - Linearity Test
  - ○ Non-linear Regression
- Model Fine Tuning
  - ○ Outlier and Heteroskedasticity Detection
  - ○ Collinearity Assessment
  - ○ Re-evaluation and Modification of Polynomial Degrees
- Prediction on Test Data

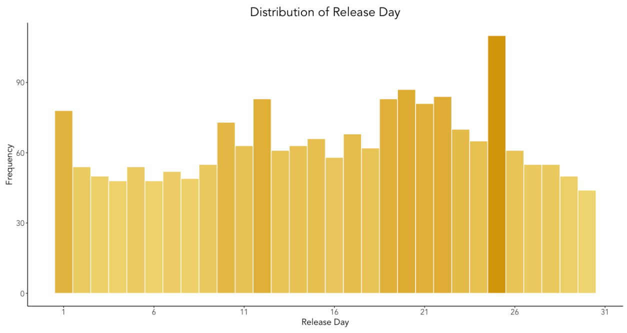
After executing the aforementioned steps, we derived a model with a R-squared of 50%. To enhance this model's precision, we suggest expanding the sample size and incorporating more pertinent predictors to the IMDb score. Conclusively, our model, built on data from around 2000 IMDb movies, effectively predicts ratings for the 12 forthcoming blockbuster films and highlights key factors influencing high IMDb scores.

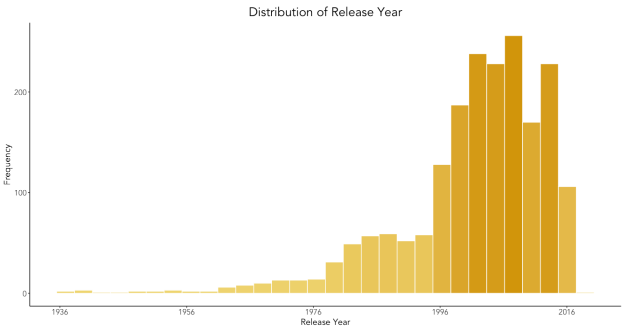
5 Appendices

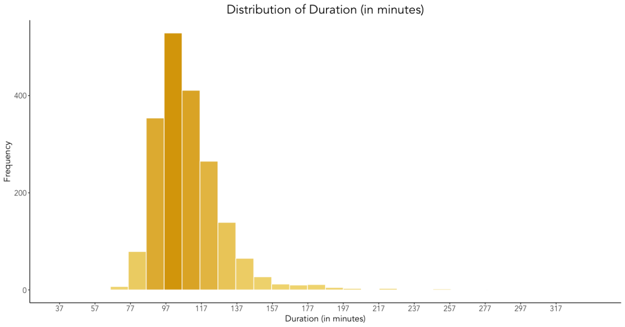
5.1 Appendix - Variable Distribution Plots

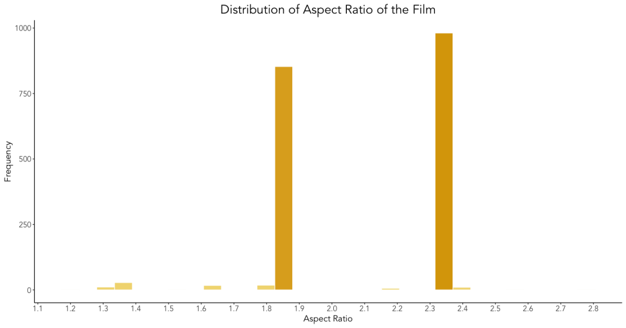


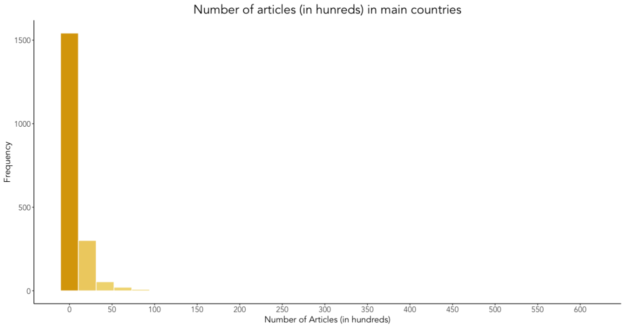


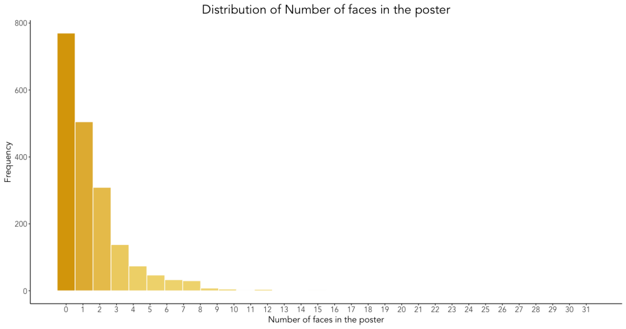


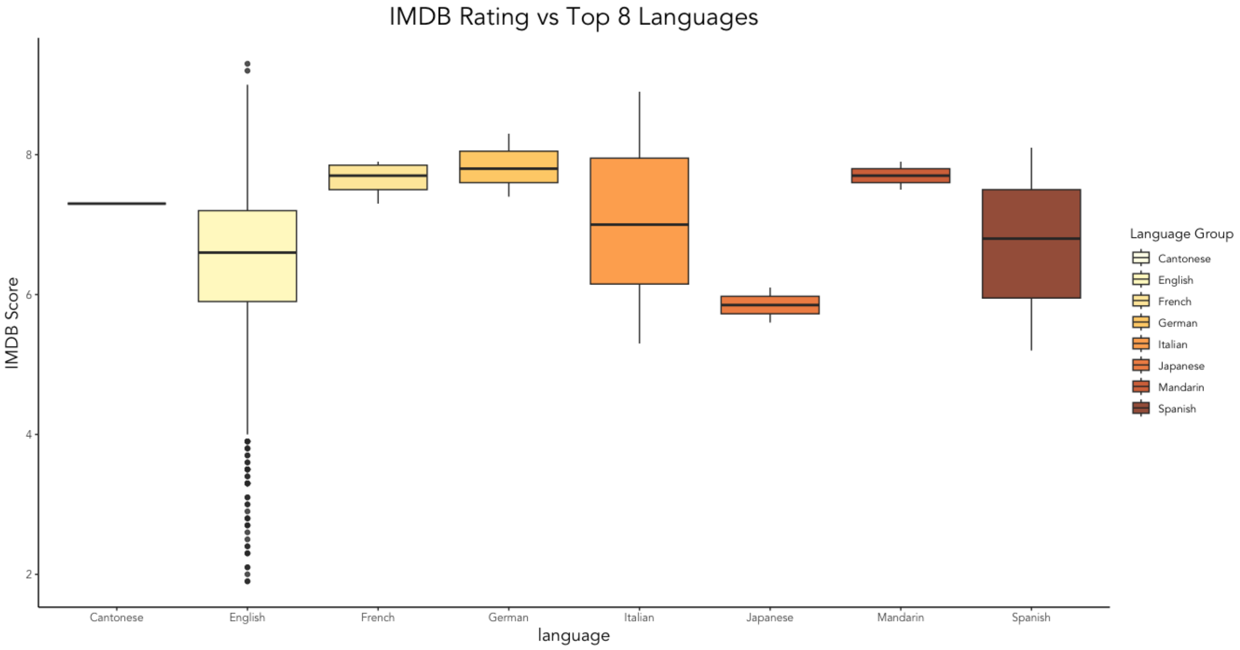


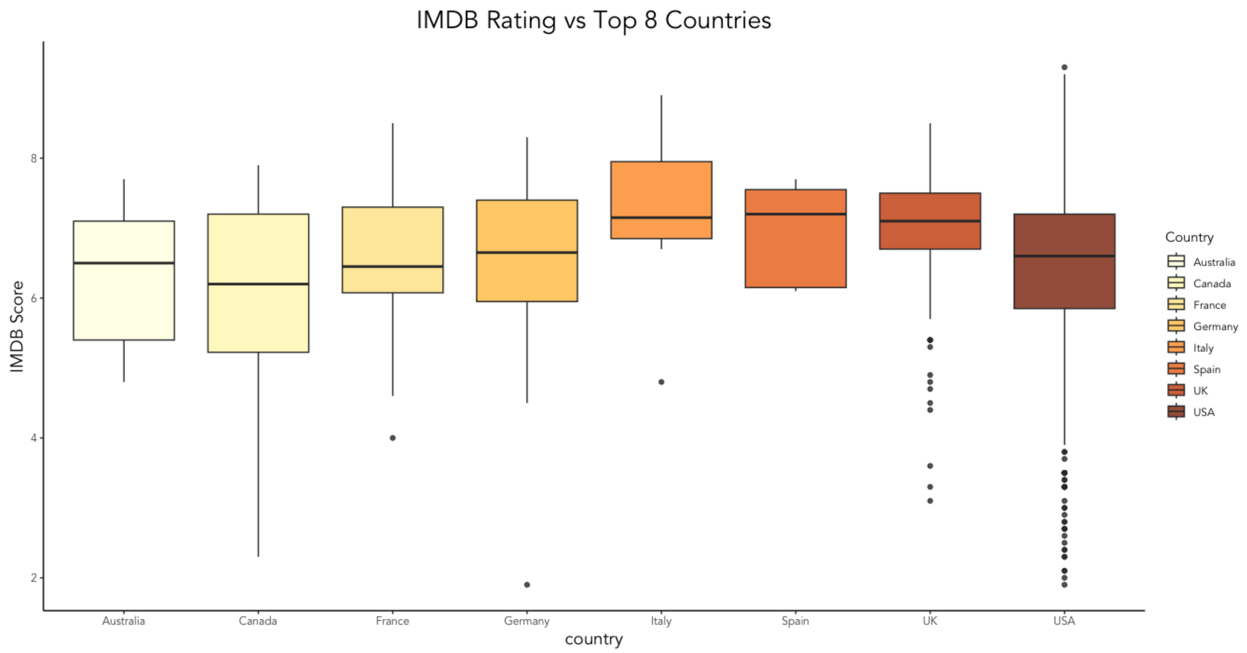

Distribution of Duration (in minutes)
Frequency
0
200
400
37
57
77
97
117
137
157
177
197
217
237
257
277
297
317
Duration (in minutes)

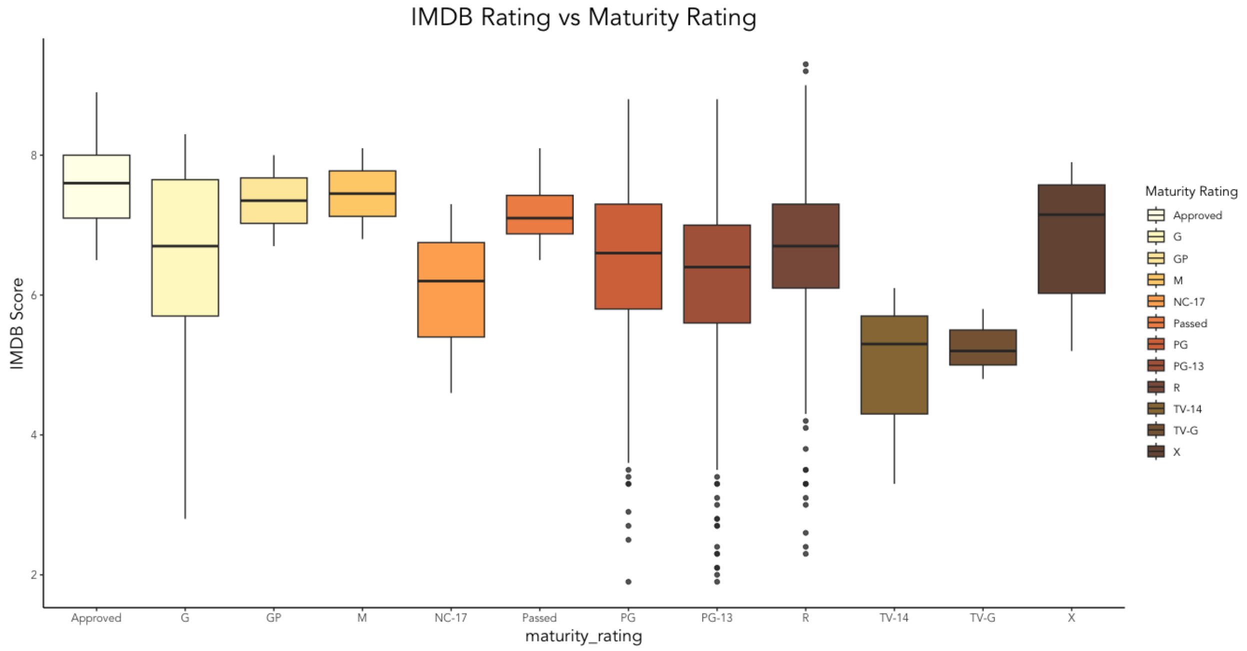

Distribution of Aspect Ratio of the Film
Frequency
0
250
500
750
1000
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1
2.2
2.3
2.4
2.5
2.6
2.7
2.8
Aspect Ratio

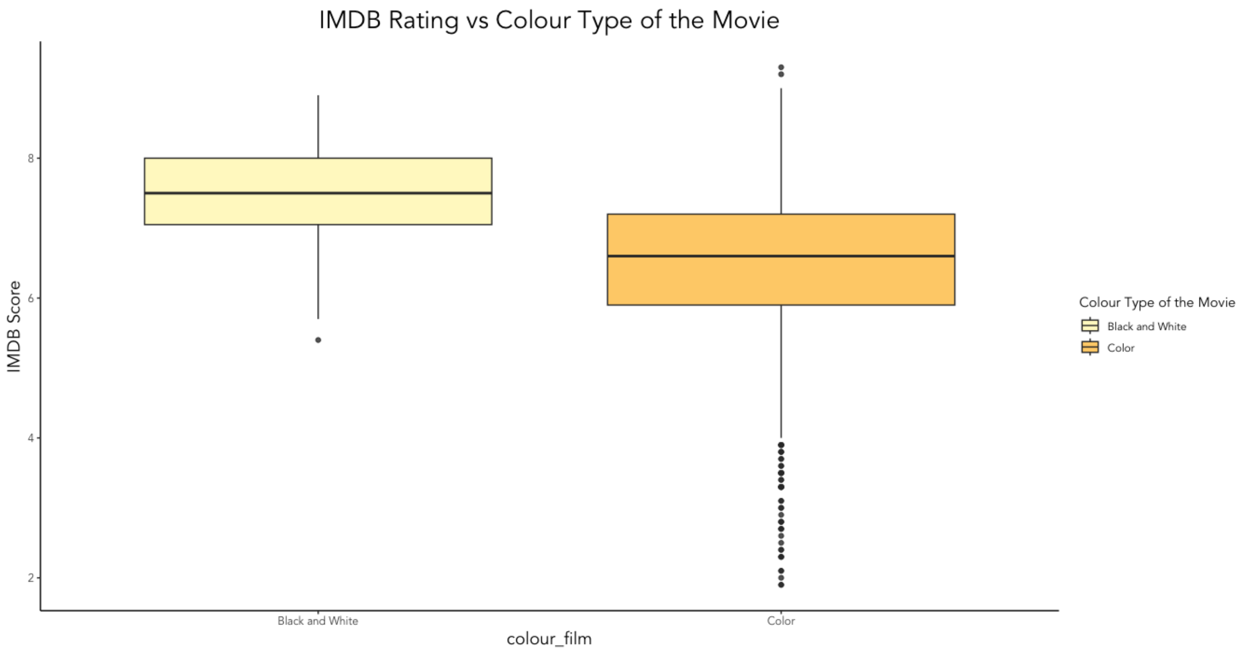

Number of articles (in hunreds) in main countries
Frequency
0
500
1000
1500
0
50
100
150
200
250
300
350
400
450
500
550
600
Number of Articles (in hundreds)

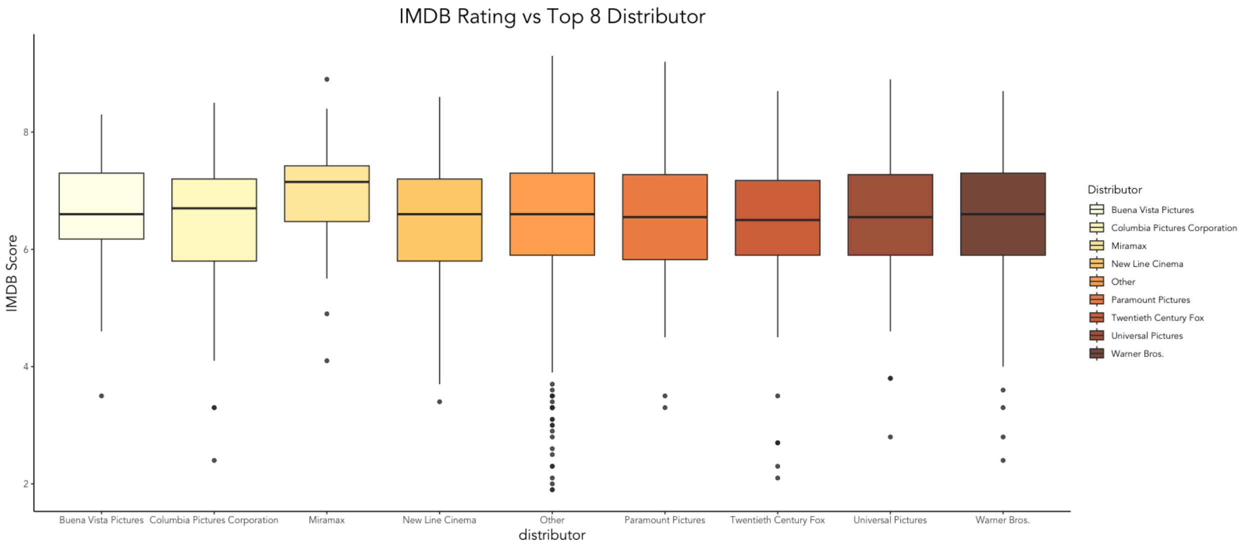

Distribution of Number of faces in the poster
Frequency
0
200
400
600
800
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
Number of faces in the poster


IMDB Rating vs Top 8 Languages
IMDB Score
2
4
6
8
Cantonese
English
French
German
Italian
Japanese
Mandarin
Spanish
language
Language Group
Cantonese
English
French
German
Italian
Japanese
Mandarin
Spanish

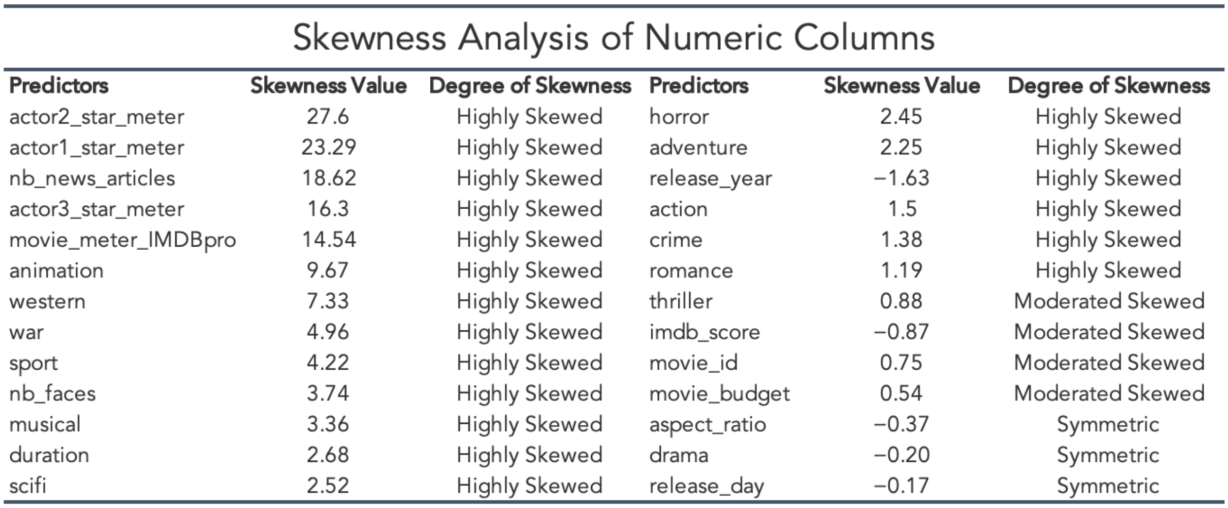

IMDB Rating vs Top 8 Countries
IMDB Score
2
4
6
8
Australia
Canada
France
Germany
Italy
Spain
UK
USA
country
Country
Australia
Canada
France
Germany
Italy
Spain
UK
USA


IMDB Rating vs Maturity Rating
IMDB Score
2
4
6
8
Approved
G
GP
M
NC-17
Passed
PG
PG-13
R
TV-14
TV-G
X
maturity_rating
Maturity Rating
Approved
G
GP
M
NC-17
Passed
PG
PG-13
R
TV-14
TV-G
X


IMDB Rating vs Colour Type of the Movie
IMDB Score
8
6
4
2
Black and White
Color
colour_film
Colour Type of the Movie
Black and White
Color


IMDB Rating vs Top 8 Distributor
IMDB Score
8
6
4
2
Buena Vista Pictures
Columbia Pictures Corporation
Miramax
New Line Cinema
Other
Paramount Pictures
Twentieth Century Fox
Universal Pictures
Warner Bros.
distributor
Distributor
Buena Vista Pictures
Columbia Pictures Corporation
Miramax
New Line Cinema
Other
Paramount Pictures
Twentieth Century Fox
Universal Pictures
Warner Bros.

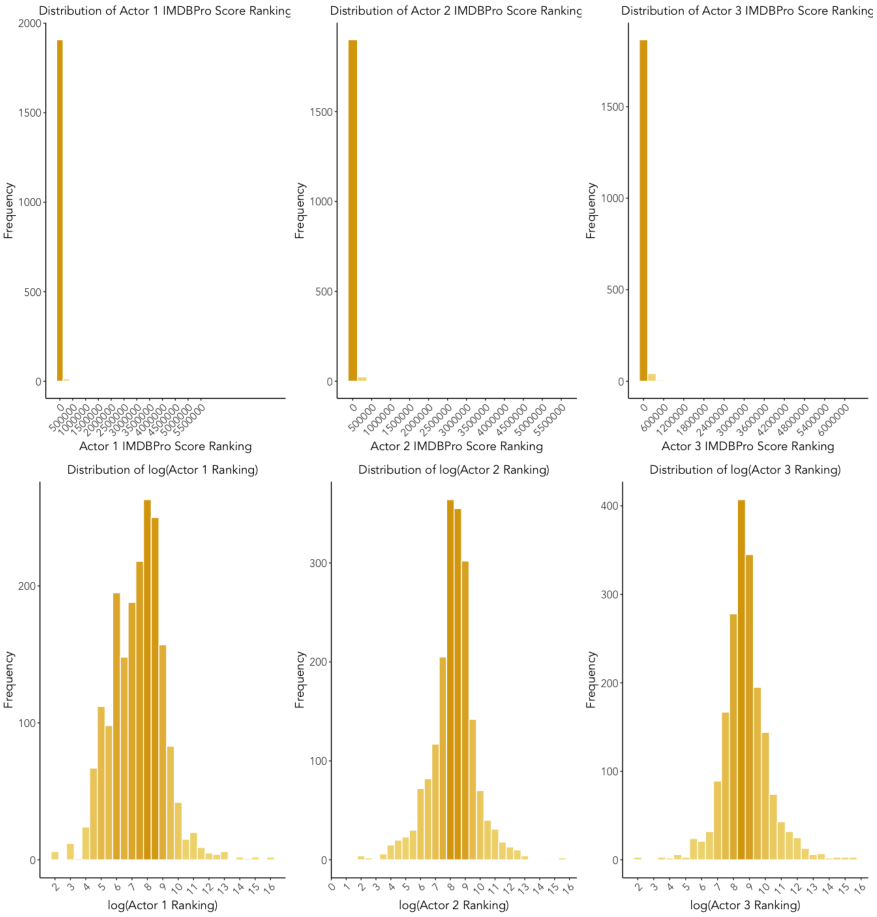
5.2 Appendix - Skewness

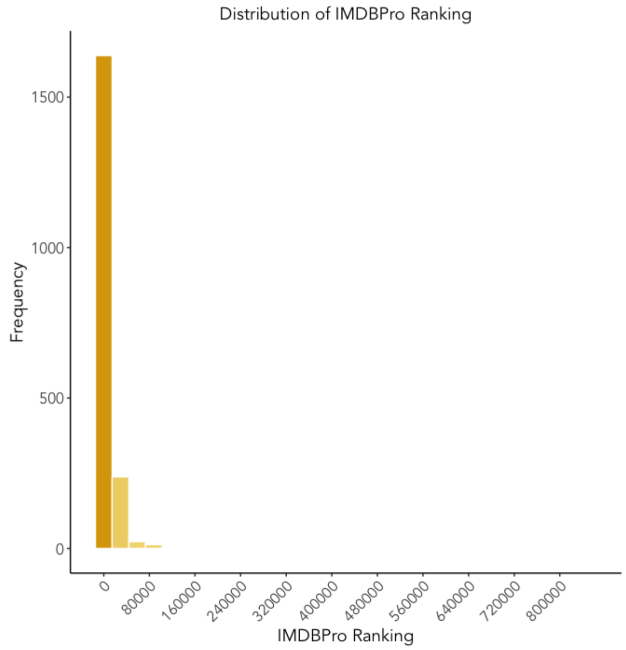


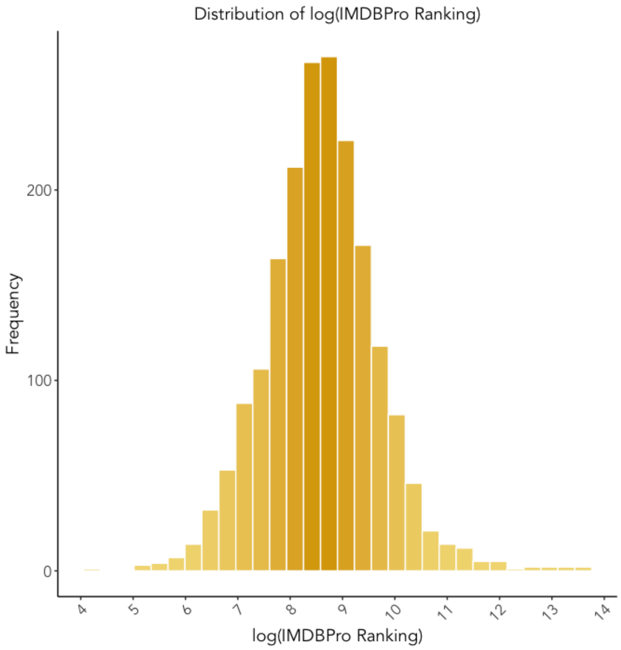
Skewness Analysis of Numeric Columns

Predictors	Skewness Value	Degree of Skewness	Predictors	Skewness Value	Degree of Skewness
actor2_star_meter	27.6	Highly Skewed	horror	2.45	Highly Skewed
actor1_star_meter	23.29	Highly Skewed	adventure	2.25	Highly Skewed
nb_news_articles	18.62	Highly Skewed	release_year	−1.63	Highly Skewed
actor3_star_meter	16.3	Highly Skewed	action	1.5	Highly Skewed
movie_meter_IMDBpro	14.54	Highly Skewed	crime	1.38	Highly Skewed
animation	9.67	Highly Skewed	romance	1.19	Highly Skewed
western	7.33	Highly Skewed	thriller	0.88	Moderated Skewed
war	4.96	Highly Skewed	imdb_score	−0.87	Moderated Skewed
sport	4.22	Highly Skewed	movie_id	0.75	Moderated Skewed
nb_faces	3.74	Highly Skewed	movie_budget	0.54	Moderated Skewed
musical	3.36	Highly Skewed	aspect_ratio	−0.37	Symmetric
duration	2.68	Highly Skewed	drama	−0.20	Symmetric
scifi	2.52	Highly Skewed	release_day	−0.17	Symmetric

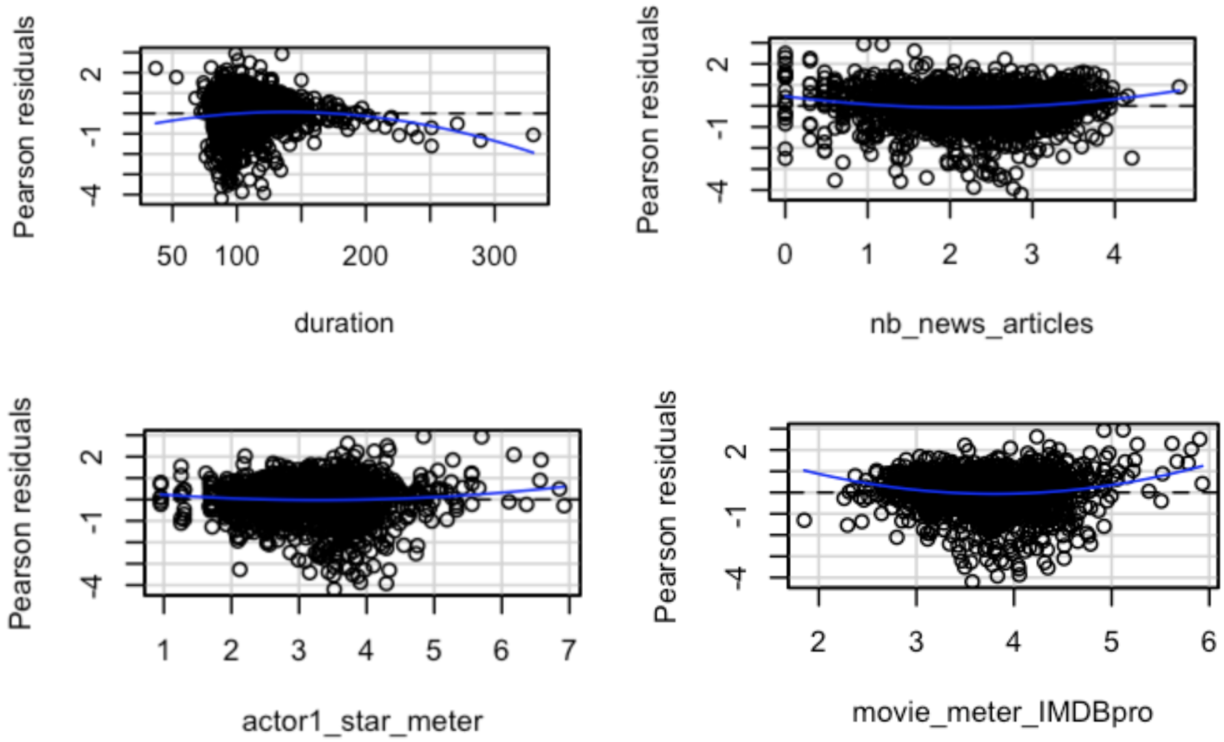
5.3 Appendix - Log-transformation



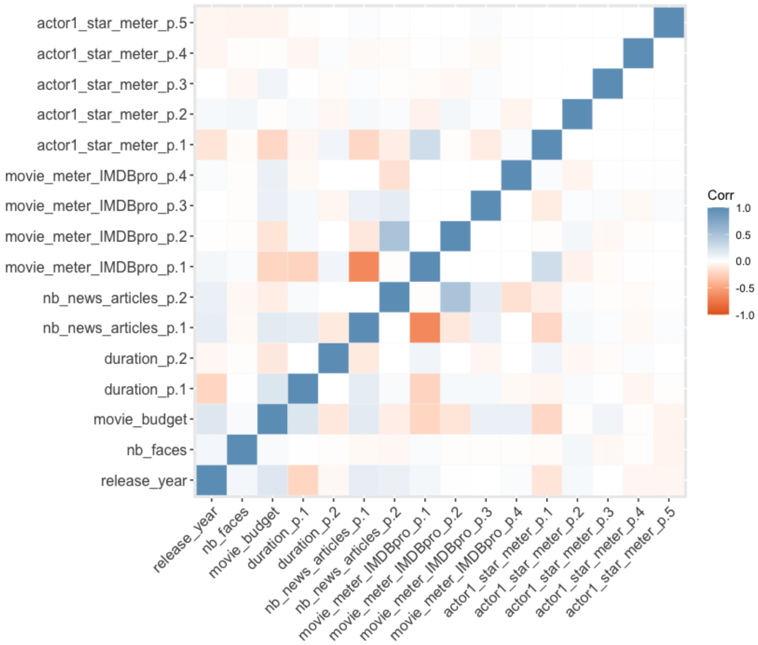




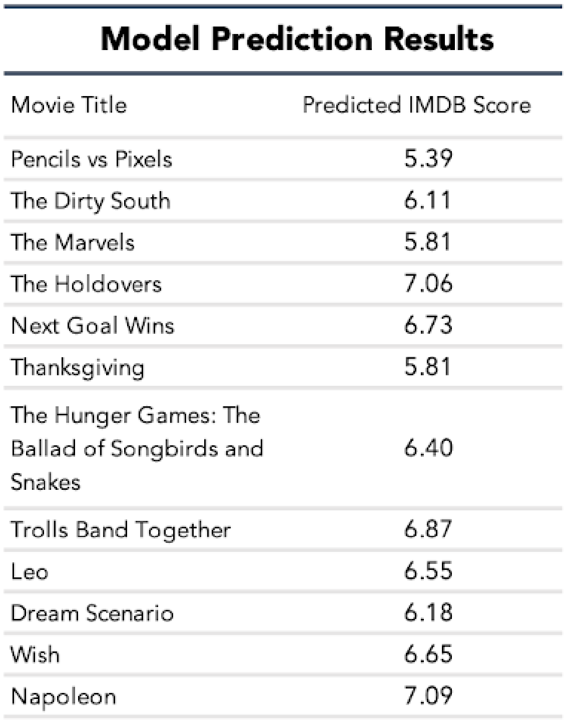
5.4 Appendix - Linearity Visual Test



5.5 Appendix - Collinearity



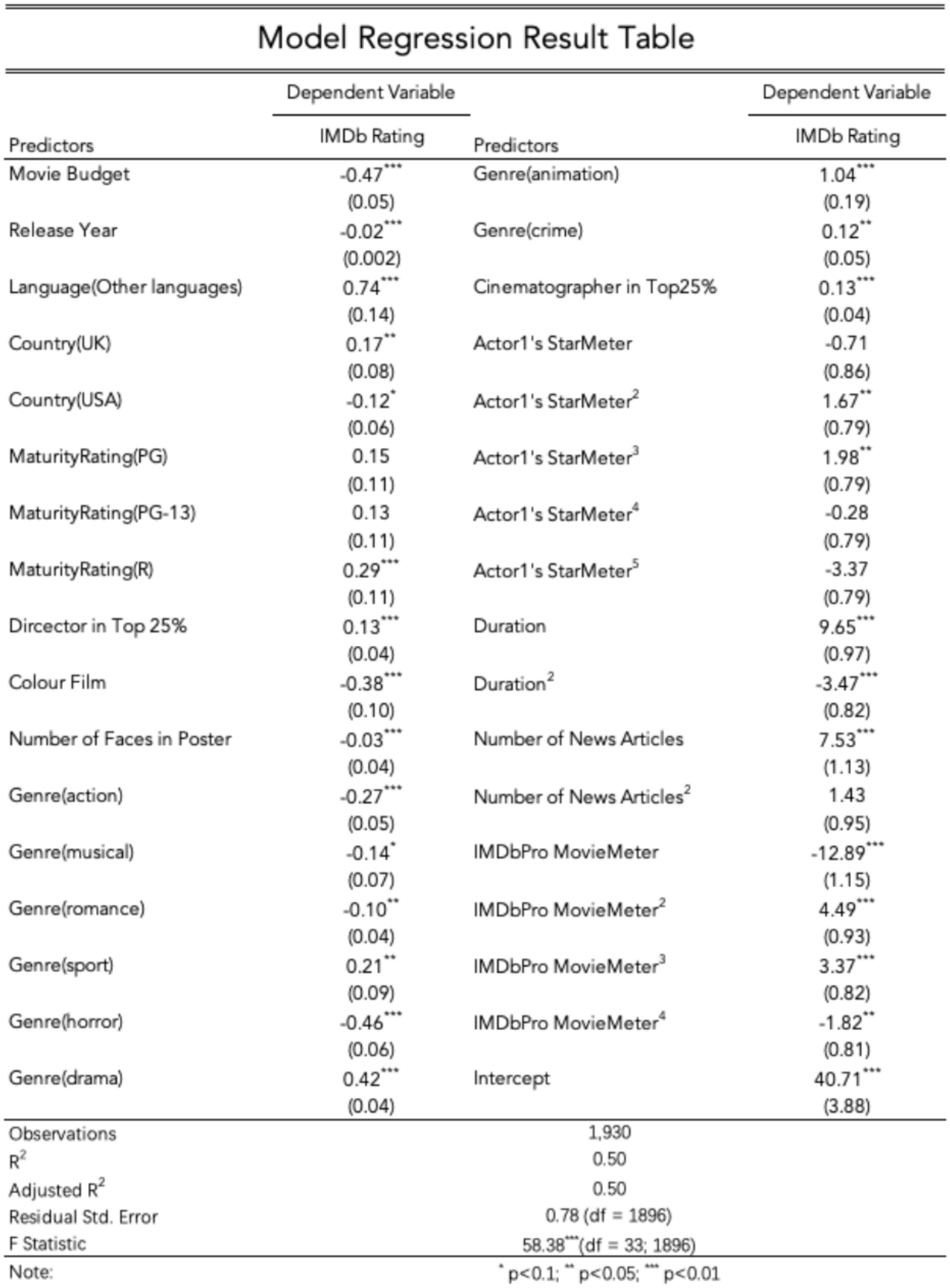
5.6 Appendix - Model Prediction Result



Model Prediction Results

Movie Title	Predicted IMDB Score
Pencils vs Pixels	5.39
The Dirty South	6.11
The Marvels	5.81
The Holdovers	7.06
Next Goal Wins	6.73
Thanksgiving	5.81
The Hunger Games: The Ballad of Songbirds and Snakes	6.40
Trolls Band Together	6.87
Leo	6.55
Dream Scenario	6.18
Wish	6.65
Napoleon	7.09

5.7 Appendix - Model Statistics



Model Regression Result Table

	Dependent Variable		Dependent Variable
Predictors	IMDb Rating	Predictors	IMDb Rating
Movie Budget	-0.47^{***}	Genre(animation)	1.04^{***}
	(0.05)		(0.19)
Release Year	-0.02^{***}	Genre(crime)	0.12^{**}
	(0.002)		(0.05)
Language(Other languages)	0.74^{***}	Cinematographer in Top25%	0.13^{***}
	(0.14)		(0.04)
Country(UK)	0.17^{**}	Actor1's StarMeter	-0.71
	(0.08)		(0.86)
Country(USA)	-0.12^{*}	Actor1's StarMeter^2	1.67^{**}
	(0.06)		(0.79)
MaturityRating(PG)	0.15	Actor1's StarMeter^3	1.98^{**}
	(0.11)		(0.79)
MaturityRating(PG-13)	0.13	Actor1's StarMeter^4	-0.28
	(0.11)		(0.79)
MaturityRating(R)	0.29^{***}	Actor1's StarMeter^5	-3.37
	(0.11)		(0.79)
Dircector in Top 25%	0.13^{***}	Duration	9.65^{***}
	(0.04)		(0.97)
Colour Film	-0.38^{***}	Duration^2	-3.47^{***}
	(0.10)		(0.82)
Number of Faces in Poster	-0.03^{***}	Number of News Articles	7.53^{***}
	(0.04)		(1.13)
Genre(action)	-0.27^{***}	Number of News Articles^2	1.43
	(0.05)		(0.95)
Genre(musical)	-0.14^{*}	IMDbPro MovieMeter	-12.89^{***}
	(0.07)		(1.15)
Genre(romance)	-0.10^{**}	IMDbPro MovieMeter^2	4.49^{***}
	(0.04)		(0.93)
Genre(sport)	0.21^{**}	IMDbPro MovieMeter^3	3.37^{***}
	(0.09)		(0.82)
Genre(horror)	-0.46^{***}	IMDbPro MovieMeter^4	-1.82^{**}
	(0.06)		(0.81)
Genre(drama)	0.42^{***}	Intercept	40.71^{***}
	(0.04)		(3.88)
Observations		1,930	
R^2		0.50	
Adjusted R^2		0.50	
Residual Std. Error		0.78 (df = 1896)	
F Statistic		58.38^{***}(df = 33; 1896)	
Note:		* $p<0.1$; ** $p<0.05$; *** $p<0.01$	