Table of Contents

1. Introduction

1. 1 Project Summary

In November 2023, there are 12 blockbuster movies being released. It is hard for filmmakers to know if their movie will be a hit and they will make back all of their initial investments. Predicting the success of these movies before their public release remains a challenging blend of art and analytics. To navigate this uncertainty, we've leveraged a dataset of 2,000 movies from IMDb. From identifiers like movie titles and IMDb links to film characteristics such as budgets, release dates, and plot keywords — and not forgetting the critical dependent variable, the IMDb score — we have a rich set of variables, giving us a well-rounded view of the films in our dataset.

1.2 Project Goals

Firstly, we aim to initially understand our dataset. By examining each variable in isolation, we hope to understand their individual distributions, identify any skewed variables, and pinpoint potential outliers that might exhibit unusual behavior. This foundational understanding will set the stage for our subsequent analyses.

Moving forward, we will look into the intricate relationships that exist within our data. By exploring the correlation between the dependent variable, IMDb ratings, and each predictor, we hope to identify patterns, strengths, and directions of these relationships. Scatter plots, non-constant variance tests, and simple linear regressions will be instrumental in this phase, shedding light on heteroskedasticity, potential collinearity, and the linear predictive power of our variables.

Once we've established this groundwork, we will build our model. Our goal is to test the linearity assumptions, explore potential non-linear relationships, and experiment with various polynomial functions and spline functional forms. With a better understanding of each predictor's relationship with our dependent variable, we'll be able to start the process of model building. This involves selecting significant predictors, considering potential interactions, and making informed decisions about the inclusion of dummy variables.

Finally, the true test of our model's success will be its out-of-sample performance. Using techniques like validation-set tests, K-fold, and LOOCV tests, we'll evaluate our model's predictive prowess, ensuring it's both robust and reliable.

In essence, our project will go through understanding, exploration, construction, and validation, with the ultimate goal of predicting the IMDb rating of the twelve upcoming blockbusters. Once the movies are released, our goal is to have the lowest MSE's compared to the true IMBD rating, hence the most accurate and reliant model.

2. Data Description

2.1 Introduction to the Dataset

We have a dataset that lists details for 1,930 movies, with each movie having 42 different pieces of information about it. This dataset was collected to help us figure out what affects IMDb movie ratings. While looking through the data, there were some data pre-processing steps that we had to take:

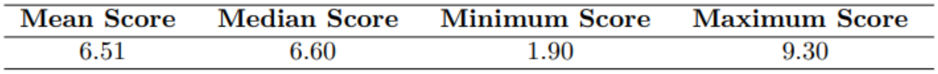
1. Drop irrelevant columns: such as movie_title, movie_id, imdb_link, as they are unique
2. Dummify categorical variables: transformed colour_film into binary and streamlined maturity_rating to four categories, converting obsolete ratings for modern compatibility.
3. Instead of dummifying actors, distributors, production companies, and directors, we categorized them by their appearance counts in our dataset, giving them a score on 1- 4.
4. Genre Adjustments: Dropped the redundant genres column, introduced missing genre columns for better clarity, and merged music into musical for consistency.

2.2 Dependent Variable (Y)

The dependent variable in our dataset is the imdb_score, which represents the IMDb rating of a movie. This score is crucial as it indicates the movie's reception and popularity among viewers.

Descriptive Statistics:

Table 1: IMDb Score Descriptive Statistics



Mean Score	Median Score	Minimum Score	Maximum Score
6.51	6.60	1.90	9.30

Visualization: Appendix A showcases the distribution in a histogram and boxplot of the IMDb scores. Most movies seem to cluster around the 6 to 7 score range, as seen through the mean and median scores.

Skewness: The IMDb scores have a slight negative skew of approximately -0.87, indicating that there are a few movies with lower ratings pulling the average down. While many movies perform within the expected range, there are certain films that did not resonate well with the audience, affecting the average.

Outliers: We utilized the Interquartile Range (IQR) method to identify potential outliers. Based on this approach, movies with IMDb scores below approximately 3.8 or above approximately 9.4 were identified as outliers. This resulted in the detection of 47 movies that fall outside these bounds. However, after careful consideration, we decided not to remove these outliers from our dataset. They are still legitimate scores between 0 and 10, hence they are still valid in our analysis.

2.3 Independent Variables x_i Overview

In our dataset, a multitude of independent variables can potentially influence IMDb scores. To simplify our analysis, we've grouped them into six main categories: Temporal, Financial, Movie

Characteristics, Production Details, Genres, and Popularity Metrics. For each group, we highlighted the groups overall relevance, key variables, as well as distribution, skewness, outliers and correlation.

Temporal Variables: Temporal factors, such as the release year, can provide insights into how cinematic preferences evolve over time. A notable variable is the release_year, which spans from 1936 to 2018. The distribution leans towards recent movies, presenting a left skew (-1.61). The predictor does show to be significant (p value<2e-16), but no significant outliers were detected. Collinearity isn't a concern in this group either.

Financial Variables: The financial aspects of movies, especially their budgets, can hint at production quality and resources. The movie_budget is a key variable, with values ranging from approximately $560,000 to $55 million. The distribution is right-skewed (0.52) with a few movies having exceptionally high budgets. The predictor is significant, with a p-score of 0.00513. There isn't a pronounced correlation between budget and IMDb score, as shown in section 2.4.

Movie Characteristics: Characteristics like movie duration and maturity rating offer insights into the movie's content and target audience. Notably, duration shows that movies are generally around the 90-130 minute mark, with a slight right skew (2.71) due to a few exceptionally long films. Its important to note that we noticed very high skewness for the actor meters predictors (17.56-38.78). They are also all not significant (p-value>than 0.05), so, we need to be wary of those predictors when choosing our model.

Production Details: Production entities play a pivotal role in a movie's quality and overall reception. The production_company variable highlights that certain companies, like "Universal Pictures", are dominant in the dataset. The distribution is diverse, with several companies present. No significant skewness is observed, but outliers exist, representing movies from lesser-known production houses that either excel or falter significantly in IMDb scores.

Genres: A movie's genre can greatly influence its reception. Some genres such as western, musical and animation and right-skewed. They are also not significant based on their p-values. Another thing to keep in mind when picking predictors is we want the genre to be present in our test set. Sci-Fi or Western are not in our testset as an example, hence will not be used.

Popularity Metrics: Metrics like IMDbPro's movie meter shed light on a movie's pre or post-release popularity. The movie_meter_IMDBpro metric, which varies widely, is a key variable. Its distribution is right-skewed (13.39), suggesting only a few movies achieve exceptionally high popularity. There's no direct strong correlation between this metric and IMDb scores. Some high-rated movies might not be initially popular, serving as outliers.

To further understand the distribution and relationships of some of the key variables in each group, please refer to Appendix B. Here, we've visualized the data of certain above-mentioned predictors using scatter plots for continuous variables and boxplots for categorical ones, providing a more tangible grasp of the dataset's nuances. In these charts we can see that some variables do have outliers and are skewed to the right/left. We will be wary of these insights when selecting our final predictors.

2.4 Correlations and Multicollinearity

We generated a heatmap to visualize the pairwise correlations. This heatmap can be found in Appendix C and provides an overview of how each variable relates to others in terms of linear correlation. Furthermore, to quantify multicollinearity concerns, we calculated the Variance Inflation Factor (VIF) for all variables. A summary of the top 10 VIF scores is available in Appendix D. A couple insights are that rating R and PG-13 show a strong negative correlation. Genres like Action, Thriller, and Crime exhibit positive correlations among themselves, and there's a notable positive relationship between movie budgets and production company popularity. However, with no variable having a Variance Inflation Factor (VIF) above 4, multicollinearity isn't a concern, eliminating the need for variable removal.

2.5 Predictive Power & Data Concerns

Several predictors were analyzed in relation to the outcome variable imdb_score, as detailed in Appendix E. The correlation coefficients range from negative to positive values, indicating varying directions of relationships. For instance, duration has a strong positive relationship with imdb_score, while release_year exhibits a strong negative association. These relationships are further emphasized by p-values, most of which are notably below the 0.05 threshold, suggesting statistical significance.

In terms of heteroskedasticity, which was evaluated using the Non-constant variance (NCV) test, predictors like movie_budget, duration, nb_news_articles, thriller, drama, crime, and movie_meter_IMDBpro displayed evidence of heteroskedasticity. This suggests that the variability of residuals is not constant across the range of these predictors.

Outliers in the predictors were identified using the Interquartile Range (IQR) method, with variables such as release_year, duration, nb_news_articles, nb_faces, and movie_meter_IMDBpro having data points that fall outside the typical range, potentially influencing the relationships observed. It's important to consider the potential impact of these outliers and heteroskedasticity on the regression model's validity and robustness.

2.6 Summary

In section 2, we assessed our dataset, highlighting the distribution of the imdb_score and the potential predictive power of variables. However, concerns such as heteroskedasticity and multicollinearity arose, which will shape our modeling decisions going forward.

3. Model Selection

3.1 Methodology

Initially, the relationship between each predictor xi and the dependent variable Y (IMDb score) was scrutinized. Recognizing that real-world relationships can take on diverse forms (not only linear), we started an exploration using polynomial models of degrees ranging from 1 to 6 for each predictor. In parallel, we also investigated spline models, adjusting both the degree (1 to 3) and the number of knots (3 to 5). This rigorous approach resulted in a total of 15 distinct models for each predictor. Our objective was to identify models that best captured the underlying relationships, as determined by the R^2 value. The top 28 models, based on the highest R^2 values, are documented in Appendix F.

3.2 Predictor Inclusion/Exclusion Rationale

In order to determine which predictors to include/exclude in our model, we looked at multiple factors as delineated in Appendix G. We analyzed the statistical relationship of each predictor with the IMDb score, focusing on metrics like correlation coefficients, p-values, and R^2 values. Furthermore, our exploration extended to polynomial and spline models, where predictors showcasing high R^2 values were given particular attention, signifying potential non-linear relationships with the IMDb score. This depth of analysis was paired with an evaluation of heteroskedasticity and the presence of outliers. Notably, the top 15 predictors, as ranked through this comprehensive process, were included in our model exploration, ensuring they contributed the most information. The rest, despite their potential relevance, were eliminated to maintain model simplicity and clarity. For instance, while predictors such as 'duration' or 'director_popularity' might demonstrate strong statistical ties, their practical significance in the cinematic realm was also considered. It's imperative to note that predictors, even with high statistical significance, could be demoted in ranking due to challenges like heteroskedasticity or outliers.

3.3 Predictor Model Rationale

Our model selection was guided by the principle of maximizing predictive power while ensuring model interpretability. Drawing insights from our preliminary analyses, we started with top-ranked predictors, like 'duration', which demonstrated a strong linear relationship with the IMDb score. We tested 18 distinct models. Each iteration represented an alteration—adding new predictors, removing non-significant ones, introducing splines and polynomials, or evaluating specific interaction terms. Guiding our decisions were key performance metrics: Mean Squared Error (MSE), R^2, and adjusted R^2. We stopped at 18 models because we didn't want to add too many predictors to avert overfitting. Based on our exploration, model 18 was selected for having the lowest MSE and the highest R^2. For a comprehensive overview of each model's testing, formula, and associated metrics, refer to Appendix H.

4. Results

4.1 Final Model

After careful exploration and adjustments, we've landed on a final model (model 18 in exploration) to predict IMDb scores for movies. This model strikes a balance between accuracy and simplicity, giving us a clear understanding of what influences movie ratings. In this section, we'll walk through its components, why certain predictors were chosen, and our predictions for the 12 upcoming movies.

Our model formula is: **imdb_score** = 34.822 - 0.759 × bs1[1](movie_meter_IMDBpro) - 0.967 × bs2(movie_meter_IMDBpro) - 1.197 × bs3(movie_meter_IMDBpro) - 1.521 × bs4(movie_meter_IMDBpro) - 1.501 × bs5(movie_meter_IMDBpro) + 0.011 × duration + 0.0004 × nb_news_articles + 0.392 × drama - 0.014 × release_year + 0.094 × director_popularity + 0.062 × actor1_popularity + 0.214 × Biography - 0.541 × horror - 0.378 × is_color + 0.226 × Rating_R - 0.271 × action - 0.329 × is_USA - 0.043 × nb_faces - 0.000 × movie_budget - 0.00000 × (duration × nb_news_articles)

In this model, we've utilized a spline transformation on movie_meter_IMDBpro to capture its non-linear relationship with IMDb scores. The interaction term between duration and nb_news_articles suggests that the combination of a movie's length and the number of news articles about it can uniquely affect its rating. Other predictors, like drama and release_year, were chosen because of their individual significance and relevance in predicting movie ratings. This model aims to provide a comprehensive yet understandable view of the factors influencing movie ratings.

4.2 Model Performance Metrics

R-squared: The R^2 value for our model stands at 0.4778, suggesting that approximately 47.78% of the variability in IMDb scores is explained by our predictors. This is a moderate value, indicating that while our model captures a significant portion of the variability in movie ratings, there's still room for improvement, potentially with the inclusion of other unexamined factors. The adjusted R^2, which penalizes the addition of non-significant predictors, is slightly lower at 0.4717. This small difference between R^2 and adjusted R^2 suggests that most predictors in the model are relevant.

Out-of-Sample Performance (Predictive Power): To ensure our model's accuracy isn't confined just to the data it was trained on, we evaluated its performance using k-fold cross-validation. We tested the model with varying values of k: 10, 20, 30, 40, and 50. Each k-fold iteration was repeated 20 times, and the average MSE for each k value was computed. The results are detailed in Appendix I. The consistency in MSE values across varying k indicates a stable model. Notably, the slight variations in

[1] bs1-bs5 : coefficients of the basis spline with knots at (2304, 4132, 6623, 11334)

MSE for different k values show that the model's performance is relatively insensitive to the choice of k, emphasizing its robustness. However, the consistently close MSE values around 0.6335 highlight that while the model is stable, there's a persistent error term that could be addressed in future iterations or with more expansive data.

4.3 Predictor Significance

The significance of each predictor in our model was determined using its p-value, which measures the likelihood that the observed effect of a predictor is due to random chance. A lower p-value suggests a statistically significant relationship with the IMDb score. For a comprehensive view of each predictor's p-value, refer to Appendix J for the final model summary statistics. Several key findings emerge from the results:

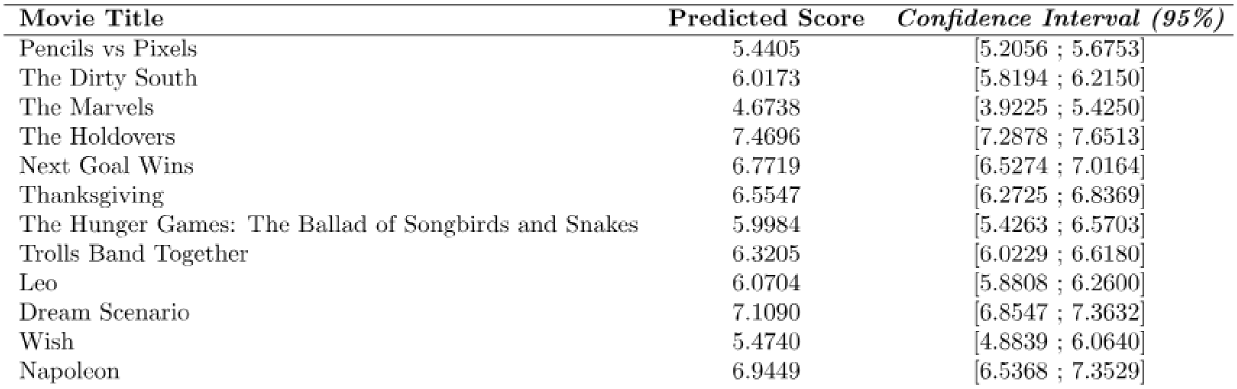
- The movie_meter_IMDBpro variable, when transformed using basis splines, showed significant non-linear effects across its range, indicating its complex relationship with the IMDb score.
- The duration of the movie held statistical significance, with both its linear and quadratic terms having an impact on the IMDb score.
- Movies classified under the drama genre were notably influential, suggesting that drama movies are more popular compared to other genres.
- While movie_budget was a statistically significant predictor, its direct influence on the IMDb score was nuanced. It’s coefficient is so low because the values for this variables are very high.
- The interaction term between duration and nb_news_articles highlighted the combined effect of movie length and media coverage on ratings.

It's vital to remember that while a low p-value (statistical significance) points to a predictor's mathematical importance in the model, its real-world impact might vary. A variable's practical significance could differ from its statistical significance, especially when taking into account real-world contexts and intricacies.

4.4 Predictions for the 12 movies

Utilizing our refined regression model, we forecasted the IMDb scores for the 12 upcoming blockbusters. This predictive task was crucial, serving as a practical application of our analytical project.

Table 10 : Predicted IMDb Scores and Confidence Intervals (95%) for test data



Movie Title	**Predicted Score**	***Confidence Interval (95%)***
Pencils vs Pixels	5.4405	[5.2056 ; 5.6753]
The Dirty South	6.0173	[5.8194 ; 6.2150]
The Marvels	4.6738	[3.9225 ; 5.4250]
The Holdovers	7.4696	[7.2878 ; 7.6513]
Next Goal Wins	6.7719	[6.5274 ; 7.0164]
Thanksgiving	6.5547	[6.2725 ; 6.8369]
The Hunger Games: The Ballad of Songbirds and Snakes	5.9984	[5.4263 ; 6.5703]
Trolls Band Together	6.3205	[6.0229 ; 6.6180]
Leo	6.0704	[5.8808 ; 6.2600]
Dream Scenario	7.1090	[6.8547 ; 7.3632]
Wish	5.4740	[4.8839 ; 6.0640]
Napoleon	6.9449	[6.5368 ; 7.3529]

4.5 Conclusion of results

It's great that we predicted these scores, but what value can they bring to a certain stakeholder looking into our analysis? Our results can have many business implications in the real world such as;

<u>Informed Decision-Making:</u> Stakeholders, ranging from producers to marketers, can leverage our model's predictions to make informed decisions. For movies predicted to score highly, marketing teams can capitalize on the positive buzz. Conversely, for those forecasted to underperform, targeted marketing strategies can be crafted to increase audience reception.

<u>Budget Allocation:</u> Predictions on movie reception can guide financial decisions, helping in the efficient allocation of budgets, especially in areas like post-production enhancements or marketing campaigns.

<u>Content Strategy:</u> Production houses can get insights into genre preferences, optimal movie durations, and other influential factors, guiding future movie projects and content strategies.

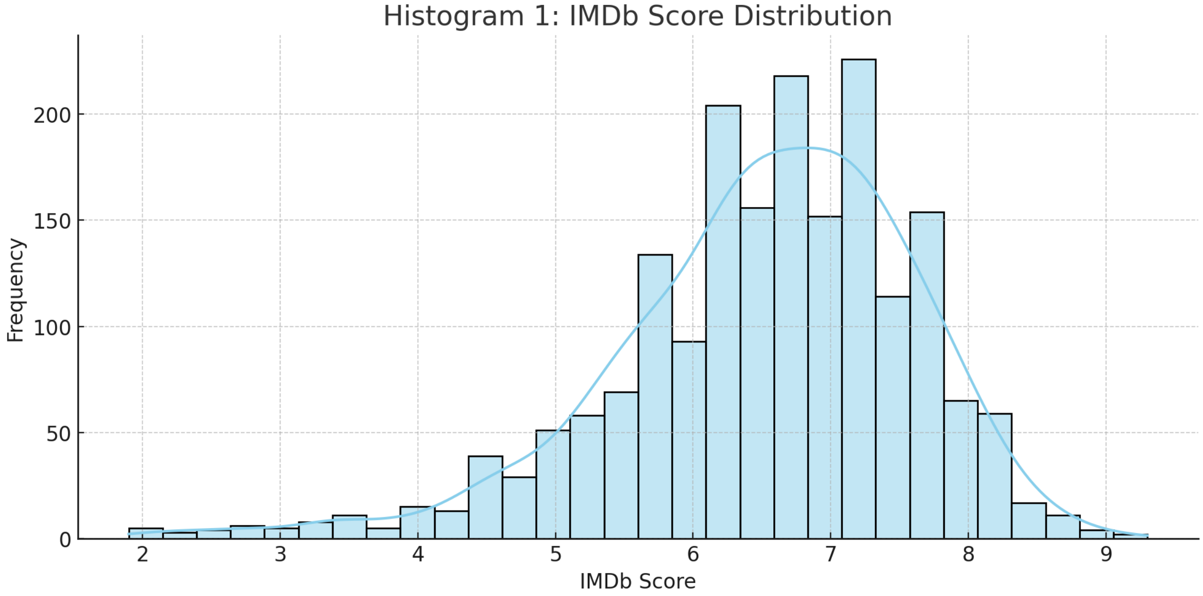
<u>Risk Management:</u> By having an early estimate of a movie's potential reception, producers can devise risk mitigation strategies, such as adjusting release dates to avoid clashes with major blockbusters or exploring alternate distribution methods.

Our regression model captures patterns in our data but might not consider all external factors or sudden industry changes. When predicting the reception of the 12 upcoming movies, we should see our forecasts as educated guesses. The film world is unpredictable; unexpected hits can arise, and big films might not always succeed. In the future, updating our model with fresh data or using advanced techniques could improve our predictions. For now however, our model shows how data can help make decisions in the movie business.

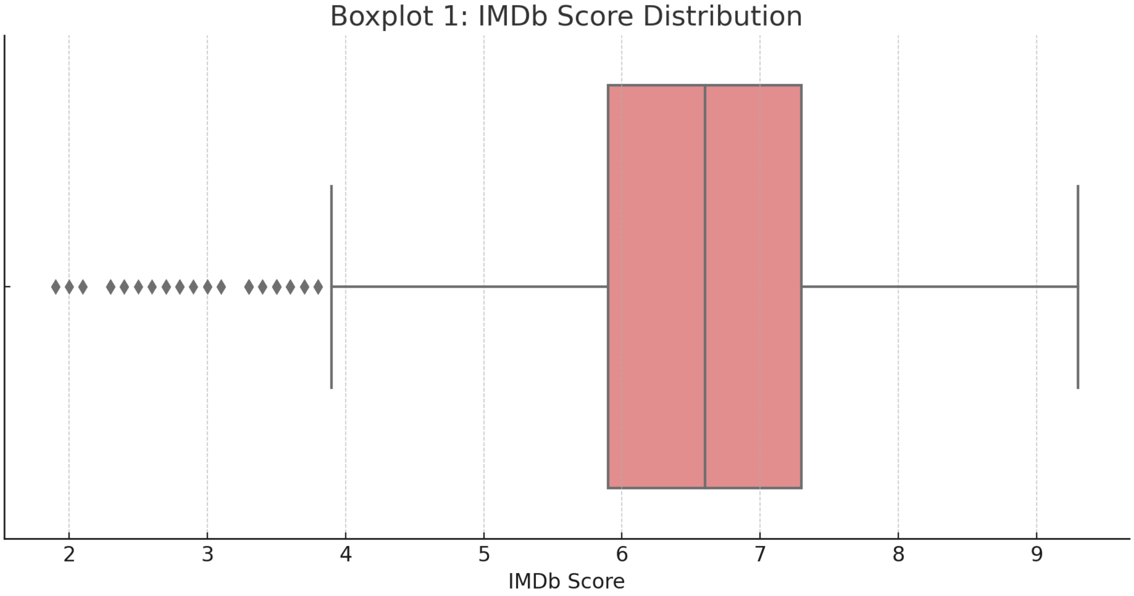
5. Appendices

Appendix A : IMDb Score Distribution

Graph 1 - IMDb Score Distribution - Histogram

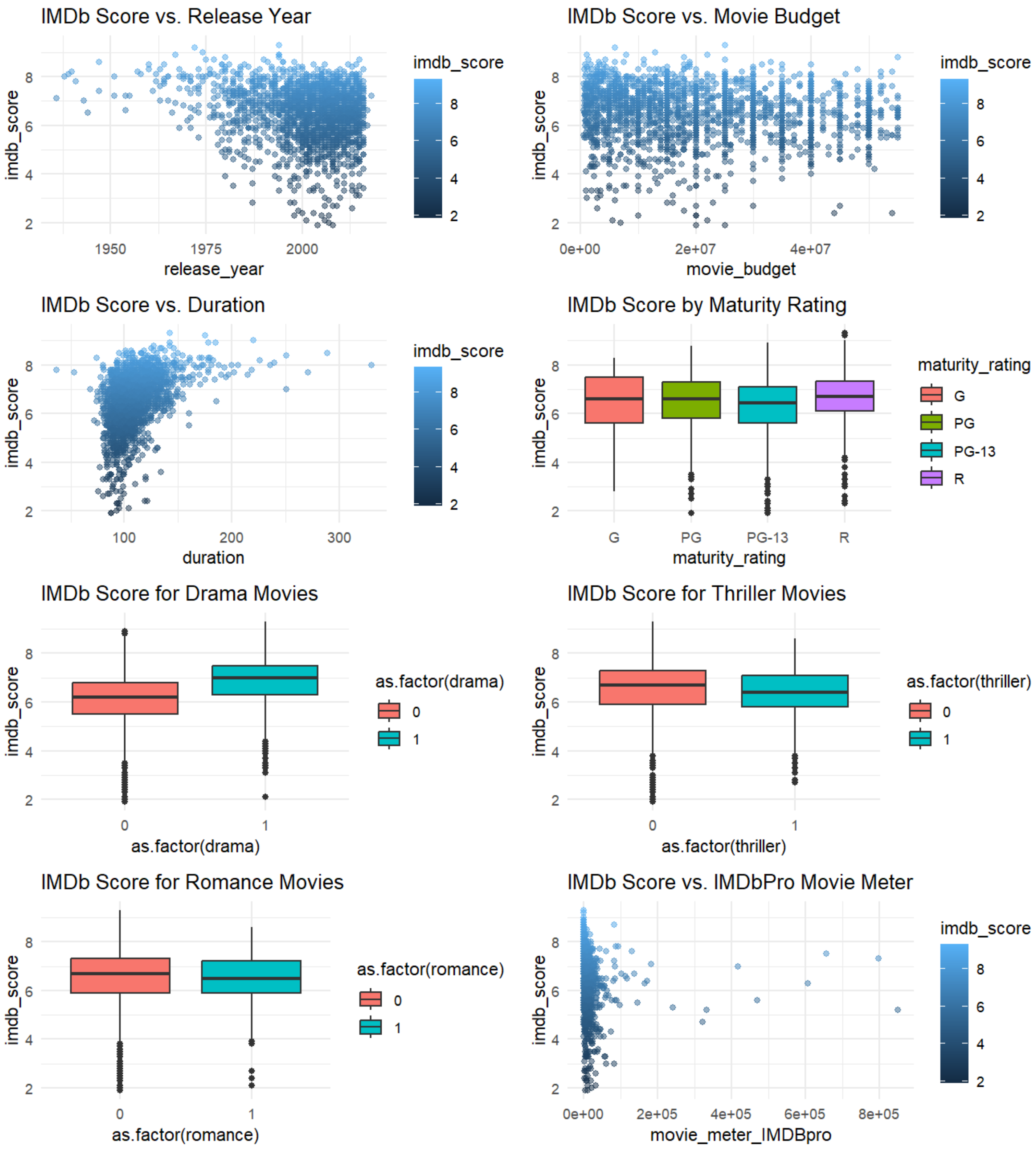


Graph 2 - IMDb Score Distribution - Boxplot



Appendix B : Variable Analysis Plots

Graphs 3-10 : Scatter Plots and Boxplots of IMDb Score vs 8 Variables



IMDb Score vs. Release Year

IMDb Score vs. Movie Budget

IMDb Score vs. Duration

IMDb Score by Maturity Rating

IMDb Score for Drama Movies

IMDb Score for Thriller Movies

IMDb Score for Romance Movies

IMDb Score vs. IMDbPro Movie Meter

Appendix C : Collinearity Heat Map & Matrix

Graph 11 : Correlation Heatmap of IMDb data

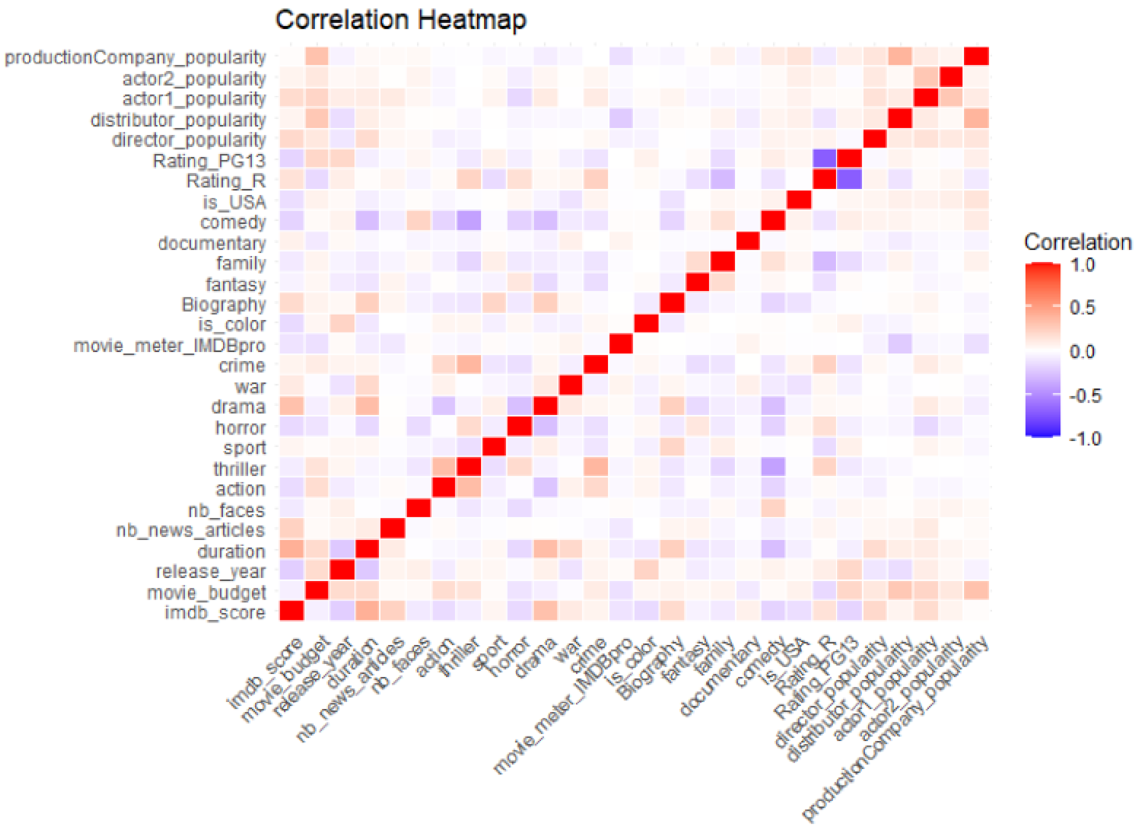
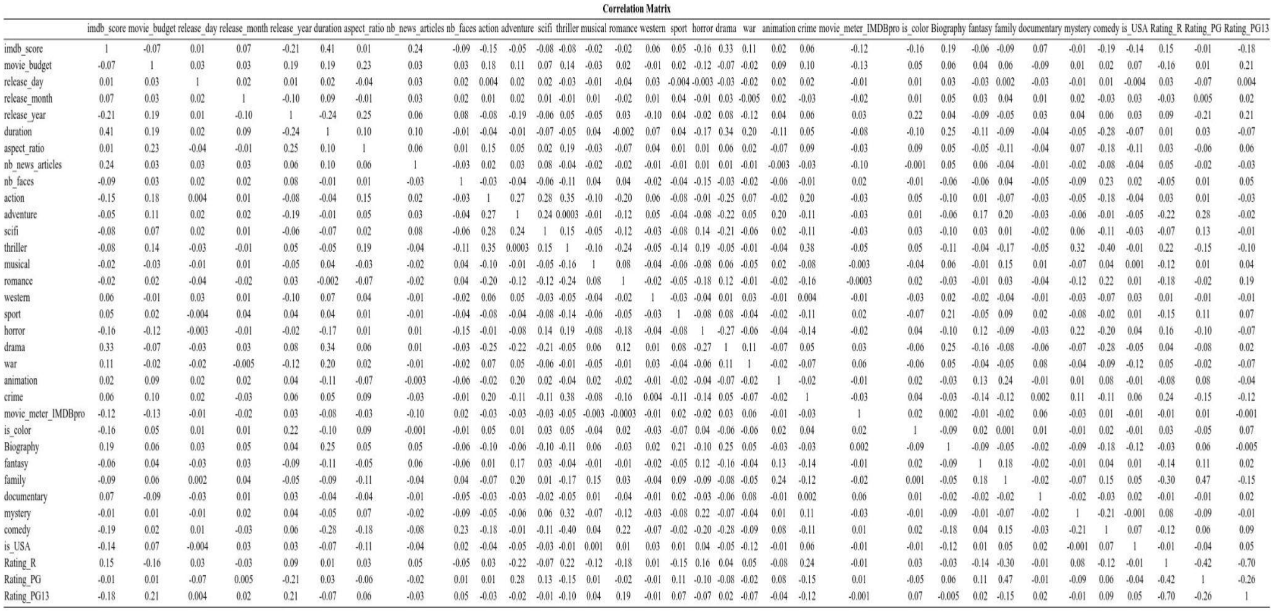


Table 2 : Correlation Matrix of IMDb data

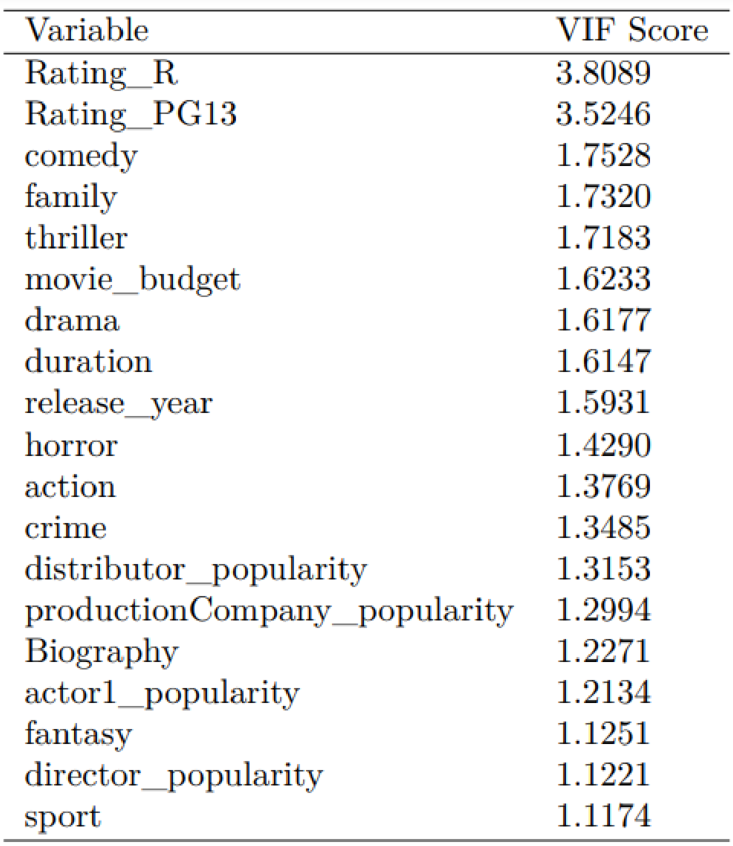


Correlation Matrix

	imdb_score	movie_budget	release_day	release_month	release_year	duration	aspect_ratio	nb_news_articles	nb_faces	action	adventure	scifi	thriller	musical	romance	western	sport	horror	drama	war	animation	crime	movie_meter_IMDBpro	is_color	Biography	fantasy	family	documentary	mystery	comedy	is_USA	Rating_R	Rating_PG	Rating_PG13
imdb_score	1	-0.07	0.01	0.07	-0.21	0.41	0.01	0.24	-0.09	-0.15	-0.05	-0.08	-0.08	-0.02	-0.02	0.06	0.05	-0.16	0.33	0.11	0.02	0.06	-0.12	-0.16	0.19	-0.06	-0.09	0.07	-0.01	-0.19	-0.14	0.15	-0.01	-0.18
movie_budget	-0.07	1	0.03	0.03	0.19	0.19	0.23	0.03	0.03	0.18	0.11	0.07	0.14	-0.03	0.02	-0.01	0.02	-0.12	-0.07	-0.02	0.09	0.10	-0.13	0.05	0.06	0.04	0.06	-0.09	0.01	0.02	0.07	-0.16	0.01	0.21
release_day	0.01	0.03	1	0.02	0.01	0.02	-0.04	0.03	0.02	0.004	0.02	0.02	-0.03	-0.01	-0.04	0.03	-0.004	-0.003	-0.03	-0.02	0.02	0.02	-0.01	0.01	0.03	-0.03	0.002	-0.03	-0.01	0.01	-0.004	0.03	-0.07	0.004
release_month	0.07	0.03	0.02	1	-0.10	0.09	-0.01	0.03	0.02	0.01	0.02	0.01	-0.01	0.01	-0.02	0.01	0.04	-0.01	0.03	-0.005	0.02	-0.03	-0.02	0.01	0.05	0.03	0.04	0.01	0.02	-0.03	0.03	-0.03	0.005	0.02
release_year	-0.21	0.19	0.01	-0.10	1	-0.24	0.25	0.06	0.08	-0.08	-0.19	-0.06	0.05	-0.05	0.03	-0.10	0.04	-0.02	0.08	-0.12	0.04	0.06	0.03	0.22	0.04	-0.09	-0.05	0.03	0.04	0.06	0.03	0.09	-0.21	0.21
duration	0.41	0.19	0.02	0.09	-0.24	1	0.10	0.10	-0.01	-0.04	-0.01	-0.07	-0.05	0.04	-0.002	0.07	0.04	-0.17	0.34	0.20	-0.11	0.05	-0.08	-0.10	0.25	-0.11	-0.09	-0.04	-0.05	-0.28	-0.07	0.01	0.03	-0.07
aspect_ratio	0.01	0.23	-0.04	-0.01	0.25	0.10	1	0.06	0.01	0.15	0.05	0.02	0.19	-0.03	-0.07	0.04	0.01	0.01	0.06	0.02	-0.07	0.09	-0.03	0.09	0.05	-0.05	-0.11	-0.04	0.07	-0.18	-0.11	0.03	-0.06	0.06
nb_news_articles	0.24	0.03	0.03	0.03	0.06	0.10	0.06	1	-0.03	0.02	0.03	0.08	-0.04	-0.02	-0.02	-0.01	-0.01	0.01	0.01	-0.01	-0.003	-0.03	-0.10	-0.001	0.05	0.06	-0.04	-0.01	-0.02	-0.08	-0.04	0.05	-0.02	-0.03
nb_faces	-0.09	0.03	0.02	0.02	0.08	-0.01	0.01	-0.03	1	-0.03	-0.04	-0.06	-0.11	0.04	0.04	-0.02	-0.04	-0.15	-0.03	-0.02	-0.06	-0.01	0.02	-0.01	-0.06	-0.06	0.04	-0.05	-0.09	0.23	0.02	-0.05	0.01	0.05
action	-0.15	0.18	0.004	0.01	-0.08	-0.04	0.15	0.02	-0.03	1	0.27	0.28	0.35	-0.10	-0.20	0.06	-0.08	-0.01	-0.25	0.07	-0.02	0.20	-0.03	0.05	-0.10	0.01	-0.07	-0.03	-0.05	-0.18	-0.04	0.03	0.01	-0.03
adventure	-0.05	0.11	0.02	0.02	-0.19	-0.01	0.05	0.03	-0.04	0.27	1	0.24	0.0003	-0.01	-0.12	0.05	-0.04	-0.08	-0.22	0.05	0.20	-0.11	-0.03	0.01	-0.06	0.17	0.20	-0.03	-0.06	-0.01	-0.05	-0.22	0.28	-0.02
scifi	-0.08	0.07	0.02	0.01	-0.06	-0.07	0.02	0.08	-0.06	0.28	0.24	1	0.15	-0.05	-0.12	-0.03	-0.08	0.14	-0.21	-0.06	0.02	-0.11	-0.03	0.03	-0.10	0.03	0.01	-0.02	0.06	-0.11	-0.03	-0.07	0.13	-0.01
thriller	-0.08	0.14	-0.03	-0.01	0.05	-0.05	0.19	-0.04	-0.11	0.35	0.0003	0.15	1	-0.16	-0.24	-0.05	-0.14	0.19	-0.05	-0.01	-0.04	0.38	-0.05	0.05	-0.11	-0.04	-0.17	-0.05	0.32	-0.40	-0.01	0.22	-0.15	-0.10
musical	-0.02	-0.03	-0.01	0.01	-0.05	0.04	-0.03	-0.02	0.04	-0.10	-0.01	-0.05	-0.16	1	0.08	-0.04	-0.06	-0.08	0.06	-0.05	0.02	-0.08	-0.003	-0.04	0.06	-0.01	0.15	0.01	-0.07	0.04	0.001	-0.12	0.01	0.04
romance	-0.02	0.02	-0.04	-0.02	0.03	-0.002	-0.07	-0.02	0.04	-0.20	-0.12	-0.12	-0.24	0.08	1	-0.02	-0.05	-0.18	0.12	-0.01	-0.02	-0.16	-0.0003	0.02	-0.03	-0.01	0.03	-0.04	-0.12	0.22	0.01	-0.18	-0.02	0.19
western	0.06	-0.01	0.03	0.01	-0.10	0.07	0.04	-0.01	-0.02	0.06	0.05	-0.03	-0.05	-0.04	-0.02	1	-0.03	-0.04	0.01	0.03	-0.01	0.004	-0.01	-0.03	0.02	-0.02	-0.04	-0.01	-0.03	-0.07	0.03	0.01	-0.01	-0.01
sport	0.05	0.02	-0.004	0.04	0.04	0.04	0.01	-0.01	-0.04	-0.08	-0.04	-0.08	-0.14	-0.06	-0.05	-0.03	1	-0.08	0.08	-0.04	-0.02	-0.11	0.02	-0.07	0.21	-0.05	0.09	0.02	-0.08	-0.02	0.01	-0.15	0.11	0.07
horror	-0.16	-0.12	-0.003	-0.01	-0.02	-0.17	0.01	0.01	-0.15	-0.01	-0.08	0.14	0.19	-0.08	-0.18	-0.04	-0.08	1	-0.27	-0.06	-0.04	-0.14	-0.02	0.04	-0.10	0.12	-0.09	-0.03	0.22	-0.20	0.04	0.16	-0.10	-0.07
drama	0.33	-0.07	-0.03	0.03	0.08	0.34	0.06	0.01	-0.03	-0.25	-0.22	-0.21	-0.05	0.06	0.12	0.01	0.08	-0.27	1	0.11	-0.07	0.05	0.03	-0.06	0.25	-0.16	-0.08	-0.06	-0.07	-0.28	-0.05	0.04	-0.08	0.02
war	0.11	-0.02	-0.02	-0.005	-0.12	0.20	0.02	-0.01	-0.02	0.07	0.05	-0.06	-0.01	-0.05	-0.01	0.03	-0.04	-0.06	0.11	1	-0.02	-0.07	0.06	-0.06	0.05	-0.04	-0.05	0.08	-0.04	-0.09	-0.12	0.05	-0.02	-0.07
animation	0.02	0.09	0.02	0.02	0.04	-0.11	-0.07	-0.003	-0.06	-0.02	0.20	0.02	-0.04	0.02	-0.02	-0.01	-0.02	-0.04	-0.07	-0.02	1	-0.02	-0.01	0.02	-0.03	0.13	0.24	-0.01	0.01	0.08	-0.01	-0.08	0.08	-0.04
crime	0.06	0.10	0.02	-0.03	0.06	0.05	0.09	-0.03	-0.01	0.20	-0.11	-0.11	0.38	-0.08	-0.16	0.004	-0.11	-0.14	0.05	-0.07	-0.02	1	-0.03	0.04	-0.03	-0.14	-0.12	0.002	0.11	-0.11	0.06	0.24	-0.15	-0.12
movie_meter_IMDBpro	-0.12	-0.13	-0.01	-0.02	0.03	-0.08	-0.03	-0.10	0.02	-0.03	-0.03	-0.03	-0.05	-0.003	-0.0003	-0.01	0.02	-0.02	0.03	0.06	-0.01	-0.03	1	0.02	0.002	-0.01	-0.02	0.06	-0.03	0.01	-0.01	-0.01	0.01	-0.001
is_color	-0.16	0.05	0.01	0.01	0.22	-0.10	0.09	-0.001	-0.01	0.05	0.01	0.03	0.05	-0.04	0.02	-0.03	-0.07	0.04	-0.06	-0.06	0.02	0.04	0.02	1	-0.09	0.02	0.001	0.01	-0.01	0.02	-0.01	0.03	-0.05	0.07
Biography	0.19	0.06	0.03	0.05	0.04	0.25	0.05	0.05	-0.06	-0.10	-0.06	-0.10	-0.11	0.06	-0.03	0.02	0.21	-0.10	0.25	0.05	-0.03	-0.03	0.002	-0.09	1	-0.09	-0.05	-0.02	-0.09	-0.18	-0.12	-0.03	0.06	-0.005
fantasy	-0.06	0.04	-0.03	0.03	-0.09	-0.11	-0.05	0.06	-0.06	0.01	0.17	0.03	-0.04	-0.01	-0.01	-0.02	-0.05	0.12	-0.16	-0.04	0.13	-0.14	-0.01	0.02	-0.09	1	0.18	-0.02	-0.01	0.04	0.01	-0.14	0.11	0.02
family	-0.09	0.06	0.002	0.04	-0.05	-0.09	-0.11	-0.04	0.04	-0.07	0.20	0.01	-0.17	0.15	0.03	-0.04	0.09	-0.09	-0.08	-0.05	0.24	-0.12	-0.02	0.001	-0.05	0.18	1	-0.02	-0.07	0.15	0.05	-0.30	0.47	-0.15
documentary	0.07	-0.09	-0.03	0.01	0.03	-0.04	-0.04	-0.01	-0.05	-0.03	-0.03	-0.02	-0.05	0.01	-0.04	-0.01	0.02	-0.03	-0.06	0.08	-0.01	0.002	0.06	0.01	-0.02	-0.02	-0.02	1	-0.02	-0.03	0.02	-0.01	-0.01	0.02
mystery	-0.01	0.01	-0.01	0.02	0.04	-0.05	0.07	-0.02	-0.09	-0.05	-0.06	0.06	0.32	-0.07	-0.12	-0.03	-0.08	0.22	-0.07	-0.04	0.01	0.11	-0.03	-0.01	-0.09	-0.01	-0.07	-0.02	1	-0.21	-0.001	0.08	-0.09	-0.01
comedy	-0.19	0.02	0.01	-0.03	0.06	-0.28	-0.18	-0.08	0.23	-0.18	-0.01	-0.11	-0.40	0.04	0.22	-0.07	-0.02	-0.20	-0.28	-0.09	0.08	-0.11	0.01	0.02	-0.18	0.04	0.15	-0.03	-0.21	1	0.07	-0.12	0.06	0.09
is_USA	-0.14	0.07	-0.004	0.03	0.03	-0.07	-0.11	-0.04	0.02	-0.04	-0.05	-0.03	-0.01	0.001	0.01	0.03	0.01	0.04	-0.05	-0.12	-0.01	0.06	-0.01	-0.01	-0.12	0.01	0.05	0.02	-0.001	0.07	1	-0.01	-0.04	0.05
Rating_R	0.15	-0.16	0.03	-0.03	0.09	0.01	0.03	0.05	-0.05	0.03	-0.22	-0.07	0.22	-0.12	-0.18	0.01	-0.15	0.16	0.04	0.05	-0.08	0.24	-0.01	0.03	-0.03	-0.14	-0.30	-0.01	0.08	-0.12	-0.01	1	-0.42	-0.70
Rating_PG	-0.01	0.01	-0.07	0.005	-0.21	0.03	-0.06	-0.02	0.01	0.01	0.28	0.13	-0.15	0.01	-0.02	-0.01	0.11	-0.10	-0.08	-0.02	0.08	-0.15	0.01	-0.05	0.06	0.11	0.47	-0.01	-0.09	0.06	-0.04	-0.42	1	-0.26
Rating_PG13	-0.18	0.21	0.004	0.02	0.21	-0.07	0.06	-0.03	0.05	-0.03	-0.02	-0.01	-0.10	0.04	0.19	-0.01	0.07	-0.07	0.02	-0.07	-0.04	-0.12	-0.001	0.07	-0.005	0.02	-0.15	0.02	-0.01	0.09	0.05	-0.70	-0.26	1

Appendix D : VIF Chart

Table 3 : Highest VIF Scores of 18 predictors ranked highest to lowest



Variable	VIF Score
Rating_R	3.8089
Rating_PG13	3.5246
comedy	1.7528
family	1.7320
thriller	1.7183
movie_budget	1.6233
drama	1.6177
duration	1.6147
release_year	1.5931
horror	1.4290
action	1.3769
crime	1.3485
distributor_popularity	1.3153
productionCompany_popularity	1.2994
Biography	1.2271
actor1_popularity	1.2134
fantasy	1.1251
director_popularity	1.1221
sport	1.1174

Appendix E : Variable Relationships with IMDb Score

Table 4 : IMDb Score Predictors against Correlation Coefficients, P-values, Heteroskedasticity and Outliers

Predictor	**Correlation Coeff**	**Conclusion**	**P-value**	**Heteroskedasticity?**	**Outliers?**
imdb_score	1.0000	N/A	N/A	N/A	N/A
movie_budget	-0.0672	negative and weak	0.00513	yes	no
release_year	-0.2101	negative and strong	<2e-16	no	yes
duration	0.4105	positive and strong	<2e-16	yes	yes
nb_news_articles	0.2358	positive and strong	<2e-16	yes	yes
nb_faces	0.0942	positive and weak	8.60E-05	no	yes
action	-0.1514	negative and moderate	2.38E-10	no	no
thriller	-0.0803	negative and weak	0.000823	yes	no
sport	0.0514	positive and weak	0.0326	no	no
horror	-0.1622	negative and moderate	1.13E-11	no	no
drama	0.3254	positive and strong	<2e-16	yes	no
war	0.1085	positive and moderate	6.02E-06	no	no
crime	0.0593	positive and weak	0.0136	yes	no
movie_meter_IMDBpro	-0.1183	negative and moderate	7.95E-07	yes	yes
is_color	-0.1582	negative and moderate	3.60E-11	no	no
Biography	0.1878	positive and moderate	3.22E-15	no	no
fantasy	-0.0562	negative and weak	0.0193	no	no
family	-0.0916	negative and weak	0.000135	no	no
documentary	0.0716	positive and weak	0.00287	no	no
comedy	-0.1866	negative and moderate	4.96E-15	no	no
is_USA	-0.1385	negative and moderate	7.10E-09	no	no
Rating_R	0.1516	positive and moderate	2.28E-10	no	no
Rating_PG13	-0.1828	negative and moderate	1.76E-14	no	no
director_popularity	0.1968	positive and moderate	<2e-16	no	no
distributor_popularity	0.0581	positive and weak	0.0156	no	no
actor1_popularity	0.1839	positive and moderate	1.21E-14	no	no
actor2_popularity	0.0595	positive and weak	0.0133	no	no
productionCompany_popularity	0.0166	positive and weak	0.489	no	no

Appendix F : Non-linearity and fit exploration

Table 5 : Exploration of splines and polynomial predictors with highest ranked R^2 values

Variable	**Model_Type**	**Degree_or_Knots**	**R_Squared**
movie_meter_IMDBpro	Spline	Degree: 2 Knots: 5	0.2037
duration	Spline	Degree: 2 Knots: 5	0.2034
nb_news_articles	Spline	Degree: 3 Knots: 5	0.1541
drama	Linear	NA	0.1054
release_year	Spline	Degree: 1 Knots: 5	0.0585
director_popularity	Polynomial	3	0.0407
actor1_popularity	Polynomial	3	0.0364
Biography	Linear	NA	0.0346
Rating_PG13	Linear	NA	0.0343
comedy	Linear	NA	0.0342
horror	Linear	NA	0.0266
is_color	Linear	NA	0.0250
Rating_R	Linear	NA	0.0232
action	Linear	NA	0.0225
is_USA	Linear	NA	0.0203
nb_faces	Polynomial	6	0.0132
war	Linear	NA	0.0118
distributor_popularity	Polynomial	3	0.0100
family	Linear	NA	0.0085
thriller	Linear	NA	0.0066
movie_budget	Polynomial	6	0.0057
movie_budget	Polynomial	6	0.0057
documentary	Linear	NA	0.0051
productionCompany_popularity	Polynomial	3	0.0041
crime	Linear	NA	0.0036
actor2_popularity	Polynomial	3	0.0036
fantasy	Linear	NA	0.0032
sport	Linear	NA	0.0024

Appendix G : Ranked Predictors

Table 6 : Predictors ranked from 1 - 20 including their rationale, R^2 value and p-value

Position	Predictor	Rationale	R 2	P-Value
1	duration	Strong correlation coefficient of 0.4105. Extremely significant p-value.Heteroskedasticity fixed. High Rˆ2 value of 0.2034 in model (Spline Degree: 2 Knots: 5).	0.2034	<2e-16
2	drama	Strong correlation coefficient of 0.3254. Extremely significant p-value. No outliers or heteroskedasticity. High Rˆ2 value of 0.1054 in model (Linear).	0.1054	<2e-16
3	nb_news_articles	Strong correlation coefficient of 0.2358. Extremely significant p-value. Heteroskedasticity fixed. High Rˆ2 value of 0.1541 in model (Spline Degree: 3 Knots: 5).	0.1541	<2e-16
4	director_popularity	Correlation coefficient of 0.1968. Extremely significant p-value. High Rˆ2 value of 0.0407 in model (Polynomial 3).	0.0407	<2e-16
5	release_year	Strong correlation coefficient of -0.2101. Extremely significant p-value. Contains outliers. High Rˆ2 value of 0.0585 in model (Spline Degree: 1 Knots: 5).	0.0585	<2e-16
6	actor1_popularity	Correlation coefficient of 0.1839. P-value of 1.21e-14. High Rˆ2 value of 0.0364 in model (Polynomial 3).	0.0364	1.21E-14
7	Biography	Correlation coefficient of 0.1878. P-value of 3.22e-15. High Rˆ2 value of 0.0346 in model (Linear).	0.0346	3.22E-15
8	comedy	Correlation coefficient of -0.1866. P-value of 4.96e-15. High Rˆ2 value of 0.0342 in model (Linear).	0.0342	4.96E-15
9	Rating_PG13	Correlation coefficient of -0.1828. P-value of 1.76e-14. High Rˆ2 value of 0.0343 in model (Linear).	0.0343	1.76E-14
10	movie_meter_IMDBpro	Correlation coefficient of -0.1183. P-value of 7.95e-07. Presence of heteroskedasticity. Contains outliers. High Rˆ2 value of 0.2037 in model (Spline Degree: 2 Knots: 5).	0.2037	7.95E-07
11	horror	Correlation coefficient of -0.1622. P-value of 1.13e-11. High Rˆ2 value of 0.0266 in model (Linear).	0.0266	1.13E-11
12	is_color	Correlation coefficient of -0.1582. P-value of 3.6e-11. High Rˆ2 value of 0.0250 in model (Linear).	0.0250	3.6E-11
13	Rating_R	Correlation coefficient of 0.1516. P-value of 2.28e-10. High Rˆ2 value of 0.0232 in model (Linear).	0.0232	2.28E-10
14	action	Correlation coefficient of -0.1514. P-value of 2.38e-10. High Rˆ2 value of 0.0225 in model (Linear).	0.0225	2.38E-10
15	is_USA	Correlation coefficient of -0.1385. P-value of 7.1e-09. High Rˆ2 value of 0.0203 in model (Linear).	0.0203	7.1E-09
16	war	Correlation coefficient of 0.1085. P-value of 6.02e-06. High Rˆ2 value of 0.0118 in model (Linear). This predictor will be automatically eliminated in the model building.	0.0118	6.02E-06
17	family	Correlation coefficient of -0.0916. P-value of 0.000135. High Rˆ2 value of 0.0085 in model (Linear). This predictor will be automatically eliminated in the model building.	0.0085	0.000135
18	nb_faces	Correlation coefficient of 0.0942. P-value of 8.6e-05. Contains outliers. High Rˆ2 value of 0.0132 in model (Polynomial 6). This predictor will be automatically eliminated in the model building.	0.0132	0.000086
19	distributor_popularity	Correlation coefficient of 0.0581. P-value of 0.0156. High Rˆ2 value of 0.0100 in model (Polynomial 3). This predictor will be automatically eliminated in the model building.	0.0100	0.0156
20	documentary	Correlation coefficient of 0.0716. P-value of 0.00287. High Rˆ2 value of 0.0051 in model (Linear). This predictor will be automatically eliminated in the model building.	0.0051	0.00287

Appendix H : Model Testing

Table 7 : Model variations that were run with their corresponding formulas, MSE, R^2 and adjusted R^2

Model	**Formula**	**MSE for K=40**	**R2**	**Adjusted R2**
Model 1 - duration, nb_news_articles, drama, release_year, and director_popularity	imdb_score ~movie_meter_IMDBpro + duration + nb_news_articles+ drama + release_year + director_popularity	0.9032	0.2974	0.295
Model 2 - Trying polynomials on movie_meter_IMDBpro, we got the highest impact with significant result with degree 7	imdb_score~poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity	0.957	0.32	0.3891
Model 3 - Adding director_popularity	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity	0.846	0.3904	0.3857
Model 4 - Adding actor1_popularity	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity + actor1_popularity	0.8127	0.3945	0.3896
Model 5 - Adding Biography	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity + actor1_popularity + Biography	1.051	0.3984	0.3931
Model 6 - Adding Rating_PG13	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity + actor1_popularity + Biography + Rating_PG13	1.019	0.4114	0.4059
Model 7 - Adding comedy (not significant so we drop it)	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity + actor1_popularity + Biography + Rating_PG13 + comedy	1.017	0.4114	0.4056
Model 8 -Adding horror	imdb_score~poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + poly(release_year,2) + director_popularity + actor1_popularity + Biography + Rating_PG13 + horror	1.188	0.4208	0.4151
Model 9 - Adding is_color	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + Rating_PG13 + horror + is_color	1.184	0.4229	0.4172
Model 10 - Adding RatingR	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + Rating_PG13 + horror + is_color + Rating_R	1.04	0.4263	0.4203
Model 11 - Adding action (Rating PG13 loses significance so we drop it)	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + Rating_PG13 + horror + is_color + Rating_R + action	0.936	0.4382	0.4319
Model 12 - Adding is_USA	imdb_score~poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA	0.7336	0.4469	0.4408
Model 13 - Adding nb_faces	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces	0.695	0.453	0.4466
Model 14 - Adding war (not significant so we drop it)	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces + war	0.719	0.4533	0.4466
Model 15 - Adding distributor_popularity	imdb_score~poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces + distributor_popularity	0.716	0.4566	0.4499
Model 16 - Adding movie_budget	imdb_score~ poly(movie_meter_IMDBpro,7) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces + distributor_popularity + movie_budget	0.657	0.4675	0.4607
Model 17 - Trying spline for movie_meter_IMDBpro (distributot_popularity loses significance so we drop it)	imdb_score~ bs(movie_meter_IMDBpro, degree=1, df=5) + duration + nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces + distributor_popularity + movie_budget	0.641	0.4735	0.4673
Model 18 - Adding interaction between duration and nb_news_articles	imdb_score~bs(movie_meter_IMDBpro, degree=1, df=5) + duration * nb_news_articles+ drama + release_year + director_popularity + actor1_popularity + Biography + horror + is_color + Rating_R + action + is_USA + nb_faces + movie_budget	0.633	0.4777	0.4716

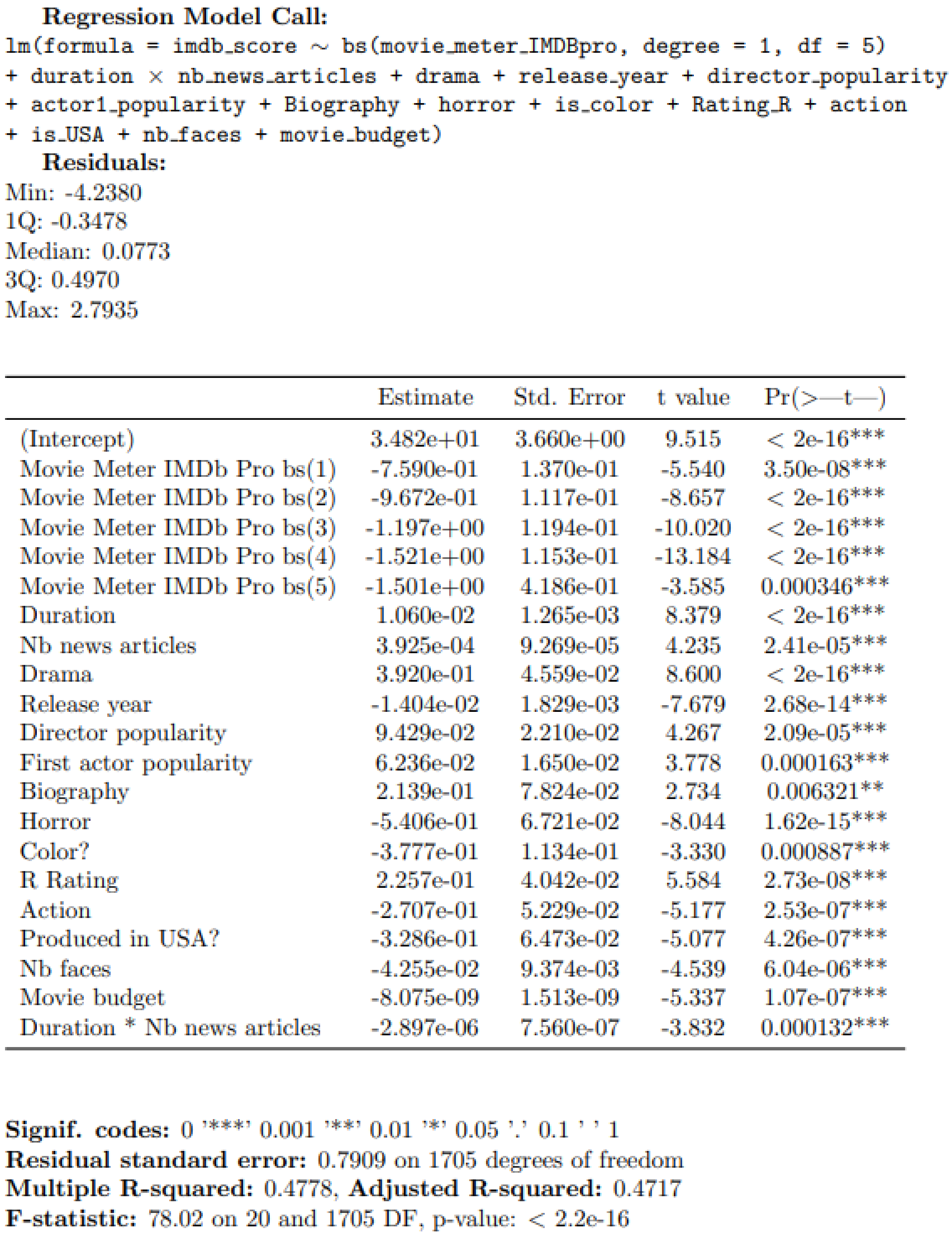
Appendix I : Out of Sample Performance

Table 8 : Average MSE of Model 18 for 5 values of K during k-fold validation test

K	**Average MSE**
10	0.6335
20	0.6341
30	0.6337
40	0.6335
50	0.6338
Total average	**0.6337**

Appendix J : Final Model Summary

Table 9 : Final Model Summary Statistics



Regression Model Call:

```
lm(formula = imdb_score ~ bs(movie_meter_IMDBpro, degree = 1, df = 5)
+ duration × nb_news_articles + drama + release_year + director_popularity
+ actor1_popularity + Biography + horror + is_color + Rating_R + action
+ is_USA + nb_faces + movie_budget)
```

Residuals:

Min: -4.2380
1Q: -0.3478
Median: 0.0773
3Q: 0.4970
Max: 2.7935

	Estimate	Std. Error	t value	Pr(>—t—)
(Intercept)	3.482e+01	3.660e+00	9.515	< 2e-16***
Movie Meter IMDb Pro bs(1)	-7.590e-01	1.370e-01	-5.540	3.50e-08***
Movie Meter IMDb Pro bs(2)	-9.672e-01	1.117e-01	-8.657	< 2e-16***
Movie Meter IMDb Pro bs(3)	-1.197e+00	1.194e-01	-10.020	< 2e-16***
Movie Meter IMDb Pro bs(4)	-1.521e+00	1.153e-01	-13.184	< 2e-16***
Movie Meter IMDb Pro bs(5)	-1.501e+00	4.186e-01	-3.585	0.000346***
Duration	1.060e-02	1.265e-03	8.379	< 2e-16***
Nb news articles	3.925e-04	9.269e-05	4.235	2.41e-05***
Drama	3.920e-01	4.559e-02	8.600	< 2e-16***
Release year	-1.404e-02	1.829e-03	-7.679	2.68e-14***
Director popularity	9.429e-02	2.210e-02	4.267	2.09e-05***
First actor popularity	6.236e-02	1.650e-02	3.778	0.000163***
Biography	2.139e-01	7.824e-02	2.734	0.006321**
Horror	-5.406e-01	6.721e-02	-8.044	1.62e-15***
Color?	-3.777e-01	1.134e-01	-3.330	0.000887***
R Rating	2.257e-01	4.042e-02	5.584	2.73e-08***
Action	-2.707e-01	5.229e-02	-5.177	2.53e-07***
Produced in USA?	-3.286e-01	6.473e-02	-5.077	4.26e-07***
Nb faces	-4.255e-02	9.374e-03	-4.539	6.04e-06***
Movie budget	-8.075e-09	1.513e-09	-5.337	1.07e-07***
Duration * Nb news articles	-2.897e-06	7.560e-07	-3.832	0.000132***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.7909 on 1705 degrees of freedom
Multiple R-squared: 0.4778, **Adjusted R-squared:** 0.4717
F-statistic: 78.02 on 20 and 1705 DF, p-value: < 2.2e-16

6. Code

Please view attached R file with this submission in order to view the code.