IMDb Prediction Challenge

Embarking on an Enchanted Journey

with "*Fairy Tail*"

Table of Contents

1. Introduction

In a world where movies are like magical tales, we, the "*Fairy Tale*" team, are embarking on a grand adventure as part of the "IMDb Prediction Challenge." We have been presented with twelve upcoming blockbusters, each with its own mysterious fate. Our mission is to predict whether these movies will be loved by the audience or met with disapproval.

Armed with our statistical tools and unwavering determination, we are setting out to create a powerful model that can foresee how well these movies will be rated on IMDb. This journey is filled with challenges and promises, much like a thrilling story. Our goal is to understand the IMDb ratings of these twelve films, like a wise oracle peering into a crystal ball to reveal their future.

We are not alone in this endeavor; we have a vast library of knowledge from over 2000 IMDb movies to guide us. With the tools of statistics, business savvy, and the spark of creativity, we navigate through complex data. As we venture forward, we encounter challenges, such as refining our predictive model. Yet, we are undaunted and believe our destiny lies in the world of IMDb.

Our adventure unfolds like a story with chapters dedicated to data exploration, variable analysis, model building, and validation. On our quest to create a powerful prediction model for the upcoming 12 blockbusters, we followed essential steps, each revealing a new part of our journey.

Each step contributes to the overall symphony of our mission to predict the fate of the 12 upcoming blockbusters:

- We began with data exploration, delving deep into the data to understand its intricacies, like diligent investigators scrutinizing variables and their relationships.
- Moving forward, we delved into variable analysis, examining data distribution, patterns, correlation, and factors influencing our focus (Y).
- We also explored the concept of non-linear relationships and considered mathematical functions and spline functions for better fit, while crafting our regression model.
- Finally, in the last chapter of our adventure, we focused on testing the model's performance with new data, diligently guarding against overfitting.

***

As the curtains rise on this magical journey, we invite you to join us on this enchanting quest.

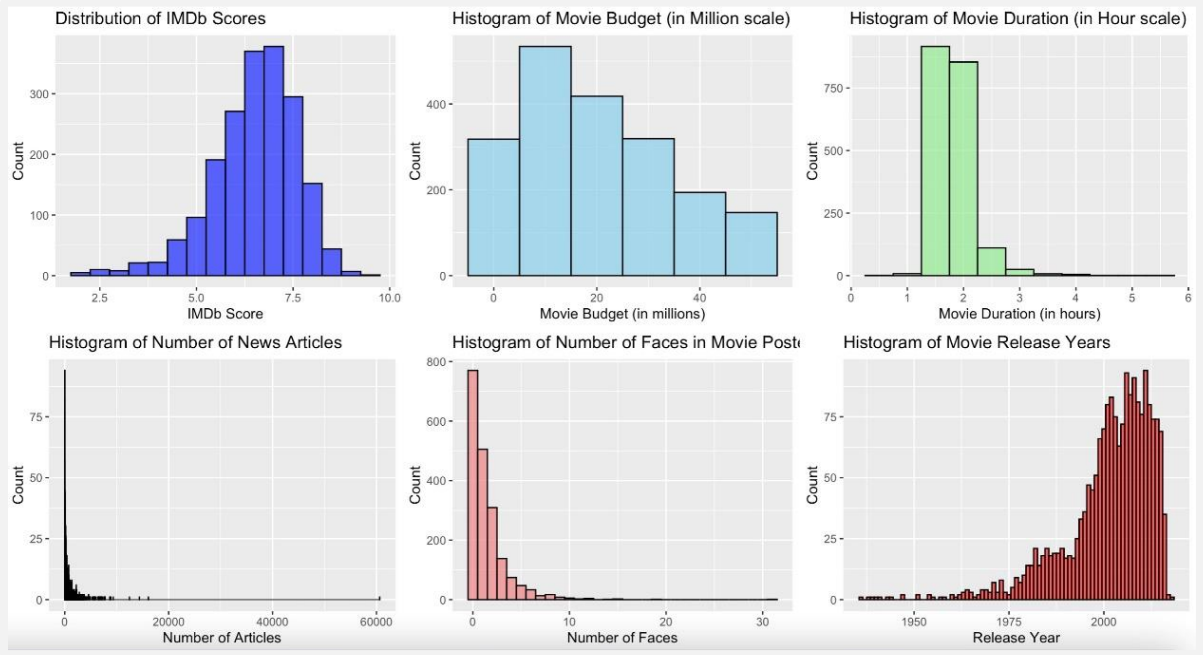
2. Data Description: Unveiling the Data's Secrets

a. Exploring Variables Individually

i. Descriptive Statistics & Distributions

Through our data analysis, several noteworthy findings have emerged. ***Figure 1*** illustrates that the distribution of IMDb scores skews to the left, signifying a predominance of movies with higher ratings as opposed to lower scores. The higher ratings indicate a potential positive bias towards well-received films by the public. Furthermore, the dataset predominantly comprises recent films with typical durations of 2 to 3 hours, which may imply that these attributes could be influential factors in determining IMDb scores. The distribution of movie budgets exhibits a right-skewed pattern, indicating a higher prevalence of movies with smaller budgets in the dataset compared to those with larger budgets. This could have implications for the model, as budget size might play a role in determining a movie's IMDb score, and it's an aspect that should be explored further in the modeling process.

Figure 1. Distribution of Numerical Variables



ii. Data Cleaning & Preprocessing

Following an in-depth examination of the dataset, we eliminated some outliers by row using qqPlot and outlierTest to improve the data's precision. Given that the data had previously been cleaned, our primary focus during preprocessing was on variable standardization. A key adjustment for this was inverting actor rankings. Originally, actors ranked as "top-rated" or “more famous” were given lower numerical values, while "low-rated" or “less famous” actors received higher ones. By inverting this system, top-rated actors now have higher numerical values, making it more intuitive when predicting IMDb scores. This adjustment ensures the influence of top-rated actors is appropriately captured in our predictive model. The inherent nature of large datasets may still result in occasional recording errors or data input inaccuracies. Minor data discrepancies are not expected to substantially affect the analysis and predictions. We

inverted the 'Movie Meter IMDb_pro' variable, which means that higher ranks now indicate a lesser impact or influence, as opposed to lower ranks signifying higher influence in the analysis.

b. Exploring Variables Relationships

i. Correlation Matrix

Subsequently, we generated a correlation matrix (depicted in Figure 2) to evaluate the relationships among the numerical variables. This would help to identify prospective predictors for our regression model where IMDb score is our target variable. The matrix unveiled specific predictors with positive correlations to the IMDb score, notably: **duration** (0.41), **drama** (0.34), **nb_news_articles** (0.23), and **war** (0.11). These insights inform our model's feature selection, as these variables are likely to play a significant role in predicting IMDb scores positively. Conversely, certain predictors demonstrated negative correlations with the score, such as **action** (-0.16), **horror** (-0.17), and **release_year** (-0.19). This understanding is crucial for our modeling process, as it suggests that these factors might negatively impact IMDb ratings.

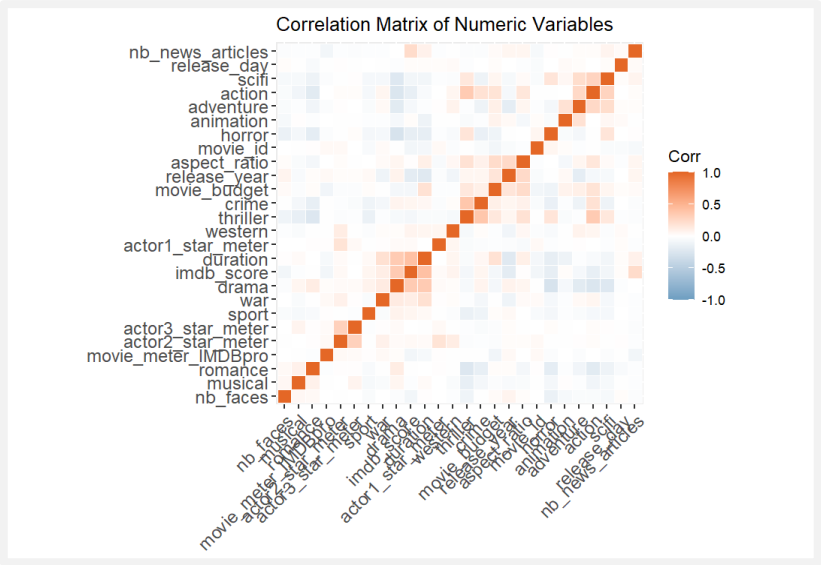


Figure 2. Correlation Matrix

Overall, these findings provide valuable guidance for selecting predictors and improving the accuracy of our IMDb score predictions.

ii. Scatterplots

When we visually examine the above correlations through scatterplots, as depicted in Figure 3, we can observe further trends. In the scatterplot of IMDb Score versus Release Year, there seems to be a dispersion of scores across various release years. No distinct linear trend is evident, suggesting that the year a movie was released may not directly influence its IMDb score. When observing IMDb Score against Duration, while there is some concentration of data points, the overall pattern does not distinctly follow a linear trend, implying that a movie's duration might not be the sole determinant of its score. Moreover, in the scatterplot representing IMDb Score versus Movie Budget, the data points are widely spread, indicating that there is not a clear linear relationship between a movie's budget and its IMDb score. This suggests that both high-budget movies and low-budget films can have varied ratings.

It is evident that these trends do not necessarily exhibit a linear pattern. This non-linearity suggests that while these predictors may have an influence on IMDb scores, they might not be the sole determinants, and their relationships with the score might be influenced by other factors or interactions. Therefore, as there may be additional variables or interactions at play, the need for non-linear modeling suggests that the predictive model should be flexible and able to accommodate these complexities.

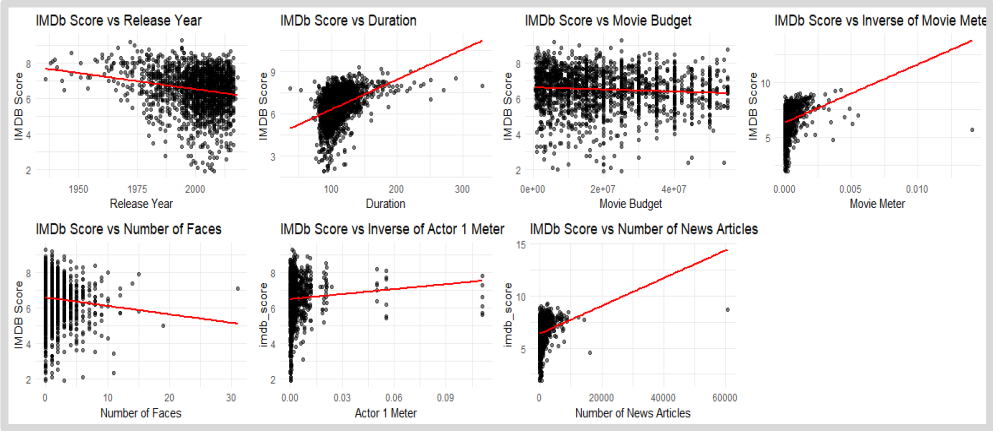


Figure 3. Scatterplots between IMDb Score vs Predictor

c. Testing Linearity & Fit

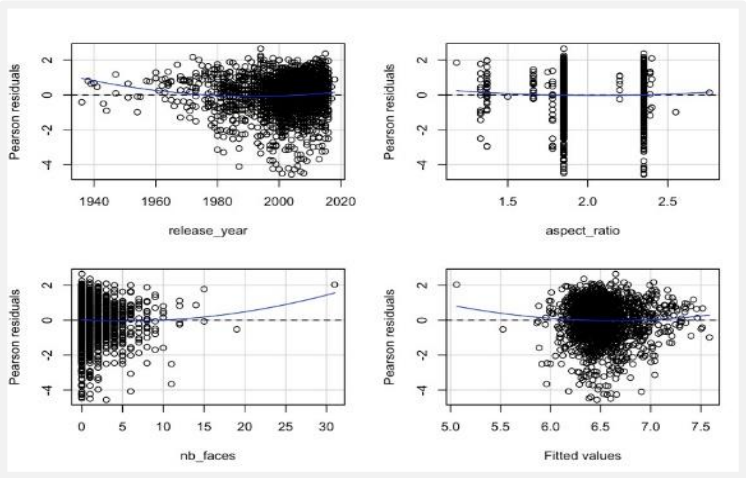
i. Simple Linear Regression

To assess whether a linear relationship exists between the IMDb score and other numerical variables, we began by fitting a linear regression using all numerical variables, with the IMDb score as the dependent variable. Figure 4 displays the residual plots for all these numerical variables, while Table 1 presents the Pearson residuals for each variable. Based on the results, for the second step of analysis, those variables with p-values lower than 0.1, including movie budget, duration, number of articles, and movie meter IMDb pro are removed. (*Cf. Appendix 3*).

ii. Collinearity

In the second step, we fitted a linear regression to the remaining numerical variables. Figure 5 displays the residual plots for all the remaining numerical variables, while Table 2 presents the Pearson residuals for those variables. We observed that the release year has a p-value less than 0.1, suggesting it should be removed, and this step should be repeated. Following this procedure, we found that even when only the variable with the highest p-value is retained, the final p-value remains less than 0.1. This indicates that there is no linear relationship between these variables and the IMDb score. (*Cf. Appendix 4*).

Figure 4. Residual Plots and Pearson Residuals of Numerical Variables



Pearson Residuals of Numerical Variables

	Dependent variable:
	IMDb Score
Release year	-0.019***
	(0.002)
Aspect ratio	0.248***
	(0.092)
Number of faces	-0.040***
	(0.012)
Intercept	44.202***
	(4.253)
Observations	1,930
R^2	0.047
Adjusted R^2	0.046
Residual Std. Error	1.075 (df = 1926)
F Statistic	31.784*** (df = 3; 1926)
Note:	*p<0.1; **p<0.05; ***p<0.01

iii. NCV Tests & Heteroskedasticity Analysis

The Non-Constant Variance (NCV) statistical test (*Appendix 5.1.*), which allows us to evaluate the presence of heteroskedasticity and examines the assumption of homoscedasticity in linear regression models, is a crucial diagnostic tool. We have addressed heteroskedasticity by incorporating inverse transformations of the actor variables in the preprocessing phase, wherein we inverted the **actor1_star_meter**, **actor2_star_meter**, **actor3_star_meter**, and **movie_meter_IMDBpro** variables. By inverting these variables, we aimed to mitigate heteroskedasticity, as lower ranks correspond to higher values.

The NCV test indicates the presence of heteroskedasticity for **movie_budget** (t-statistic = 2.4711; p-value = 0.01356 – *Cf. Appendix 5.2.*). This means that the variability of IMDb scores is not consistent across different levels of **movie_budget**. In other words, as the movie budget varies, the spread of IMDb scores also changes (inconstant level of variance for model's residuals). To address this issue, transforming the **movie_budget** variable or considering alternative modeling techniques, may be necessary to mitigate implications for the reliability of our model and to forecast the scores across a broader range of movie characteristics (*Appendix 6*).

Performing an NCV test on the final model (**reg17**) yielded a p-value of 2.22e-16 which is much less than 0.05 suggesting a strong evidence of heteroskedasticity.

iv. Modeling Nonlinear Relationships

In our exploratory analysis, we examined the relationships between the IMDb score as the dependent variable and the other variables as independents. No linear relationship was evident, as confirmed by both qualitative methods, primarily visualization, and quantitative analyses. Evidence of this non-linearity has been previously established and is documented in earlier sections through detailed visualizations and numerical data.

v. Exploring Polynomial & Spline Functions

To analyze potential non-linear relationships, we incorporated a loop within our code to evaluate polynomial regressions from degrees 2 to 6, assessing the IMDb score against each independent variable. Initially, non-numeric variables and certain genres — which functioned akin to dummy variables and displayed low p-values — were excluded. This step sharpened our focus on the polynomial regressions, enriching our data analysis. The adequacy of the polynomial models was gauged using the R-squared value as a measure of fit. Our findings revealed a pronounced polynomial relationship of the sixth degree between the IMDb score as the dependent variable and several independent variables, including 'release year,' 'duration,' 'movie budget,' 'number of articles in the news,' and 'number of faces on the main movie poster.' Consequently, these variables have been prioritized in our subsequent modeling efforts.

3. Model Selection: Choosing the Magic Formula

In this section, we delve into the model construction process, elucidating the reasoning behind modelling choices, the inclusion or exclusion of predictors, and the management of model-related issues. We initially adopted a bottom-up approach to build our model. The first iteration scrutinized the interplay between IMDb scores and movie budgets. Subsequently, the second model expanded its scope to encompass movie durations. During this initial phase, we identified an adverse relationship between IMDb scores and movie budgets. To address this discrepancy, we suspected inflation and introduced a 2% inflation rate as of the base year and standardized it. Surprisingly, the issue persisted, with movie budgets exhibiting a significant and illogical negative impact on IMDb scores and making the results unreasonable. Consequently, we decided to remove the budget from consideration.

a. Model's predictors

i. Identifying Relevant & Powerful Predictors

Our modelling journey continued by integrating categorical variables related to movie genres and release dates into the model (reg3). However, upon conducting a p-value analysis, we observed a lack of a substantial relationship between release day and release month with IMDb scores. In light of this, we opted to streamline our model by eliminating these variables.

As we advanced, we contemplated the inclusion of actor star meters (1, 2, 3) and movie meters. Our initial attempt to incorporate these variables yielded a counterintuitive result: a positive coefficient for rankings. In essence, it implied that lower rankings were associated with higher IMDb scores. To address this non-linearity, we inverted the variables to star meters and IMDb Pro ratings, resulting in a more faithful representation of the data (reg5). Subsequent analysis led to the removal of insignificant variables, further simplifying the model (reg6).

In our pursuit of model refinement, we identified and removed outliers to enhance predictive accuracy (reg7 and reg8). Furthermore, a meticulous examination of multicollinearity among numerical variables revealed no significant issues. The Variance Inflation Factor (VIF) test affirmed the absence of multicollinearity vif(reg8) (*Cf. Appendix 5.2.*). Notably, we identified a large number of unique directors relative to the dataset's size, raising concerns about potential overfitting in directors, cinematographers, production company, and distributors.

We tried to save the budget one last time by removing everything before1995, but still no help as the predictions still made no sense. To mitigate this, we split the year into 6 parts by quartile (0.2, 0.4,0.6, 0.8), which separate the old movie from the new movie and segmented them.

Our quest for model optimization involved exploring different polynomial degrees for key variables. The ultimate model (reg14) incorporated polynomial terms for release year, duration, inverted movie meter, news articles, and face counts. However, when testing the model with a validation dataset, it occasionally produced scores exceeding 10, which was implausible (thus why we didn't bother with ANOVA test). To rectify this, we seamlessly integrated IMDb's inverse movie meter with the release year and re-introduced relevant genres, effectively increasing the R-squared value and enhancing model performance (*Cf. Appendix 7*).

Here is our final model:

Reg17= **lm(formula = imdb_score ~ bs(release_year, knots = c(quantiles),degree = 1) + duration + bs(inv_movie_meter, knots = c(quantiles1), degree = 1) + nb_news_articles + nb_faces+ action + adventure + thriller + musical + romance + sport + horror + drama + war + animation + crime, data = IMDb2)**

ii. Variables Interactions & Dummies

Since there are some new director names in the test data, we had to construct a new variable, director_dummy, to model the directors. The rationale behind this choice is to categorize directors based on the number of movies they have directed, which can potentially be a significant predictor for certain outcomes. We categorized directors into quartiles based on the count of movies they directed. This categorization allows us to capture the diversity in director experience and potentially account for their influence on movie-related outcomes. After running the linear regression model, we found that adding director dummies will not improve the imdb_score accuracy. Therefore, for simplifying the model we decided to remove it. (*Cf. Appendix 8*).

4. Results: The Enchanted Revelations

a. Model's Accuracy & Predictive Performance

After seeing all the predictions of the different models made, we settled on reg15 for the resulting predictions made the most sense (at least for us). The predictions for the 12 movies in the given test dataset are as follows below:

Predicted Movie Scores

Pencils vs Pixels	The Dirty South	The Marvels	The Holdovers	Next Goal Wins	Thanksgiving	The Hunger Games: The Ballad of Songbirds and Snakes	Trolls Band Together	Leo	Dream Scenario	Wish Napoleon
5.40	5.50	6.12	6.82	6.51	7.05	7.85	6.85	6.10	6.57	8.25 7.00

	Dependent variable:
	IMDb Score
Release Year Quantile 1	-1.242*** (0.210)
Release Year Quantile 2	-1.343*** (0.185)
Release Year Quantile 3	-1.292*** (0.190)
Release Year Quantile 4	-1.377*** (0.190)
Release Year Quantile 5	-1.773*** (0.218)
Duration	0.008*** (0.001)
Inverse movie meter Quantile 1	0.291** (0.127)
Inverse movie meter Quantile 2	0.437*** (0.103)
Inverse movie meter Quantile 3	0.759*** (0.109)
Inverse movie meter Quantile 4	1.129*** (0.107)
Inverse movie meter Quantile 5	2.545*** (0.304)
Number of News Articles	0.00005*** (0.00001)
Number of Faces	-0.040*** (0.009)
Action	-0.303*** (0.054)
Adventure	-0.199*** (0.062)
Thriller	-0.101** (0.049)
Musical	-0.163** (0.075)
Romance	-0.196*** (0.047)
Sport	0.179** (0.091)
Horror	-0.462*** (0.068)
Drama	0.517*** (0.044)
War	0.221** (0.104)
Animation	0.740*** (0.190)
Crime	0.133** (0.052)
Constant	6.167*** (0.239)
Observations	1,923
R2	0.435
Adjusted R2	0.427
Residual Std. Error	0.815 (df = 1898)
F Statistic	60.772*** (df = 24; 1898)
Note:	*p<0.1; **p<0.05; ***p<0.01

The linear regression **reg 17** model, the R-squared value was around 0.43 after counting in all the significant predictors for the models and splicing it to increase it.

In the *stargazer* graph presented above, the regression summary reveals that all predictors exhibit a significance level below 0.05, indicating their statistical significance within the lm() model we've applied. Consequently, we excluded the predictors that were deemed insignificant, including two of the dummy variables representing genres (sci-fi and western). Furthermore, despite the initial significance of the 'movie budget,' it was omitted from the model due to its adverse impact on the results, which led to implausible outcomes, such as a Marvel movie receiving a rating of 4, which is unrealistic.

The slopes as well and their effects on the movie rating with a positive slope having a positive effect on the movie rating and the negative slope having a negative effect on the scope. Surprisingly the negative effect of the release year as the years progress can be explained by more bad movies being released. Also, we can see that the higher the movie meter rank, the more the movie has a chance to have a higher IMDB score. Also from the slopes, we can see that some movie genres tend to reduce the movie rating like music, and some have a positive impact such as animation. From the graphs on the top right, can still see that the general regression is still non-linear, so we still have room for improvement in the future.

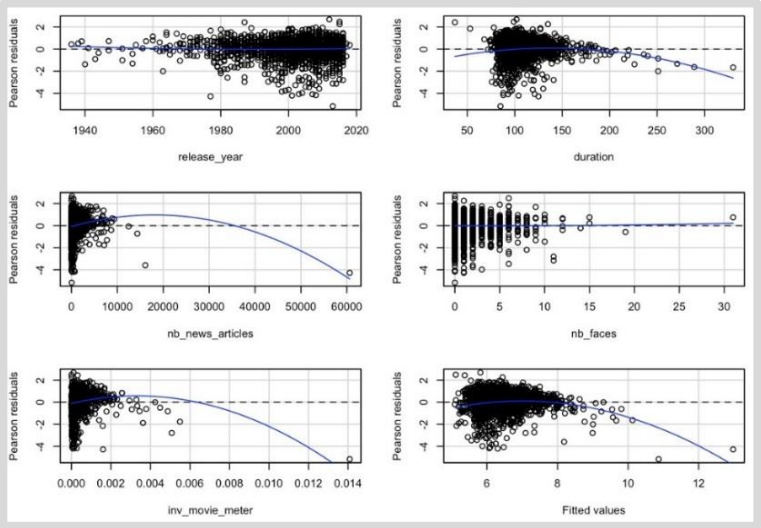


Figure 5. Pearson Residuals of Numerical Variables & Fitted Values

b. Cross-Validation Testing

To evaluate the model's predictive performance, a validation test was conducted. The dataset was divided into two sets, with a 50% split ratio. The 'train set' was employed for model training, while the 'test set' assessed its accuracy. The Mean Squared Error (MSE) value of 0.68 was computed, signifying the model's predictive accuracy.

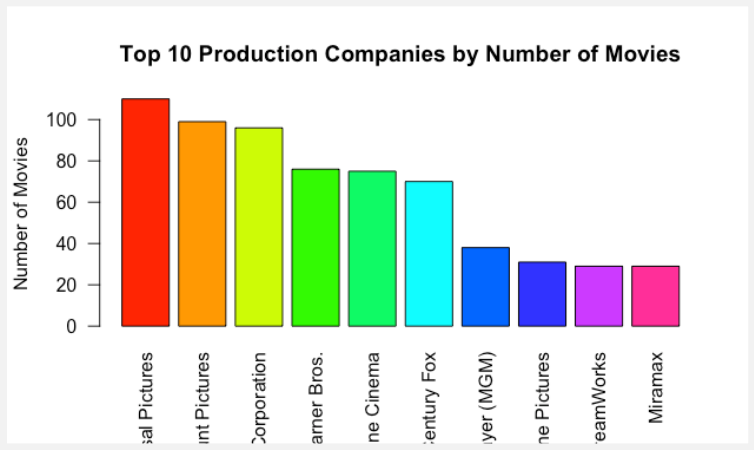
Additionally, a K-fold cross-validation with a factor of 10 was carried out, yielding an MSE of 0.68. This result indicates that, on average, the model's IMDb score predictions deviated by 0.68 from the actual scores assigned to the movies. Moreover, a Leave-One-Out Cross-Validation (**LOOCV**) was performed, also producing an MSE of 0.68. This outcome aligns with the results of the initial two tests, further underscoring the model's consistency in predicting IMDb scores.

****

In our cinematic journey to predict the fate of twelve upcoming blockbusters, we've reached our goal. Our model, though not perfect, converges toward realistic results. We've successfully solved the mystery, adding our chapter to the story of our adventurous prediction.

5. Appendix

Appendix 1: Distribution of Movies Among the Top 10 Production Companies



This figure displays the distribution of movies among the top 10 production companies, providing insights into how these companies are responsible for contributing to the film industry. It visually represents the share of movies produced by each of these leading companies.

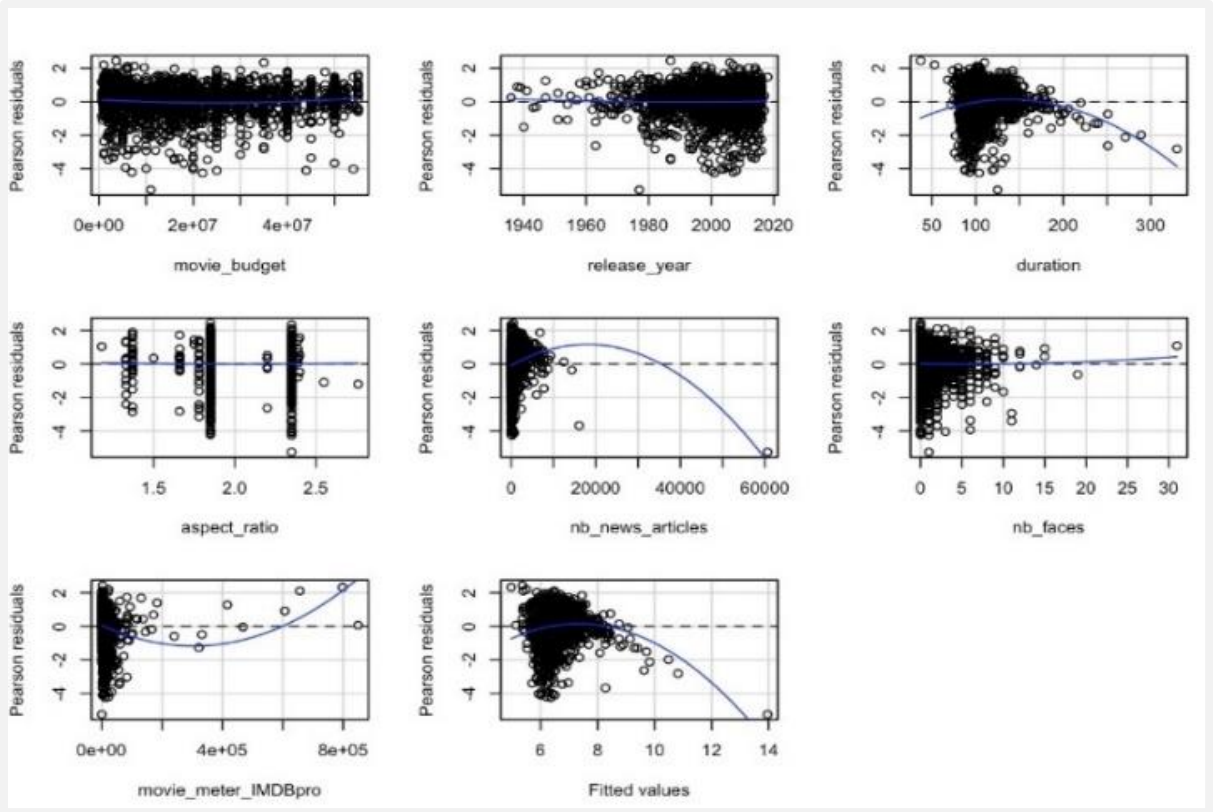
Appendix 2: Residual Plots of Numerical Variables

Pearson Residuals of Numerical Variables

	Dependent variable:
	IMDb Score
Movie budget	-0.000***
	(0.000)
Release year	-0.008***
	(0.002)
Duration	0.021***
	(0.001)
Aspect ratio	0.079
	(0.084)
Number of articles	0.0001***
	(0.00001)
Number of faces	-0.041***
	(0.011)
Movie meter IMDb pro	-0.00000***
	(0.00000)
Intercept	21.020***
	(4.016)
Observations	1,930
R^2	0.248
Adjusted R^2	0.245
Residual Std. Error	0.956 (df = 1922)
F Statistic	90.662*** (df = 7; 1922)
Note:	*p<0.1; **p<0.05; ***p<0.01

This figure showcases residual plots for a set of numerical variables, including movie budget, release year, duration, aspect ratio, number of articles, number of faces, Movie Meter IMDb Pro, and the intercept. These plots visually depict the differences between observed and predicted values for each variable, providing insights into the performance and predictive accuracy of a model or analysis related to these factors in the context of movies.

Appendix 3: Simple Linear Regressions

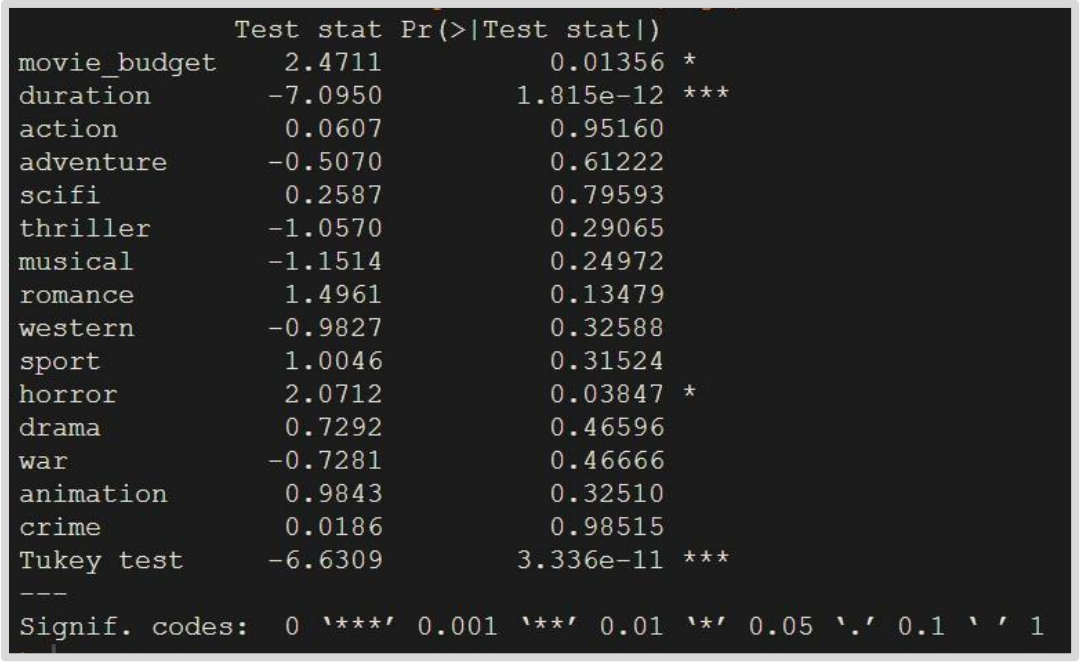


Appendix 4: Variable Inflation Factors (VIF) for Multicollinearity Assessment

```
> vif(reg8)
    movie_budget         duration       inv_actor1  inv_movie_meter nb_news_articles         nb_faces           action
        1.163307         1.335528         1.037376         1.211267         1.141062         1.060331         1.422898
       adventure            scifi         thriller          musical          romance          western            sport
        1.244369         1.223249         1.474288         1.053369         1.173532         1.031100         1.083378
          horror            drama              war        animation            crime
        1.341619         1.410233         1.100350         1.077641         1.371243
>
> # Same thing no multicollinearity
```

This figure illustrates the Variable Inflation Factors (VIF) for assessing multicollinearity within our potential model. VIF measures the degree of correlation between predictor variables, helping us identify and understand the potential issues of multicollinearity in our analysis. It provides valuable insights into the interrelationships among variables.

Appendix 5.1.: Heteroskedasticity Assessment via NCV Test

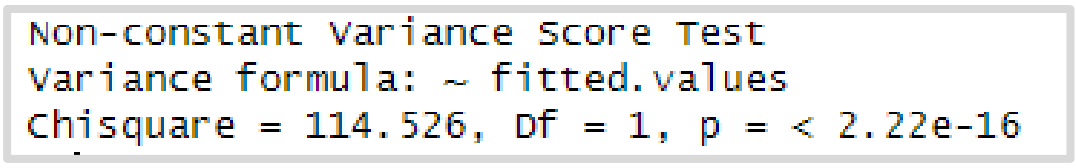


	Test stat	Pr(>|Test stat|)	
movie_budget	2.4711	0.01356	*
duration	-7.0950	1.815e-12	***
action	0.0607	0.95160	
adventure	-0.5070	0.61222	
scifi	0.2587	0.79593	
thriller	-1.0570	0.29065	
musical	-1.1514	0.24972	
romance	1.4961	0.13479	
western	-0.9827	0.32588	
sport	1.0046	0.31524	
horror	2.0712	0.03847	*
drama	0.7292	0.46596	
war	-0.7281	0.46666	
animation	0.9843	0.32510	
crime	0.0186	0.98515	
Tukey test	-6.6309	3.336e-11	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

This figure presents the results of a heteroskedasticity assessment conducted using the NCV (Non-Constant Variance) test. The predictors included in this test are 'movie_budget,' 'duration,' and a combination of 'genres' that encompass various movie genres. The figure aims to evaluate whether there is evidence of non-constant variance in the model and provides insights into the potential impact of heteroskedasticity on the analysis.

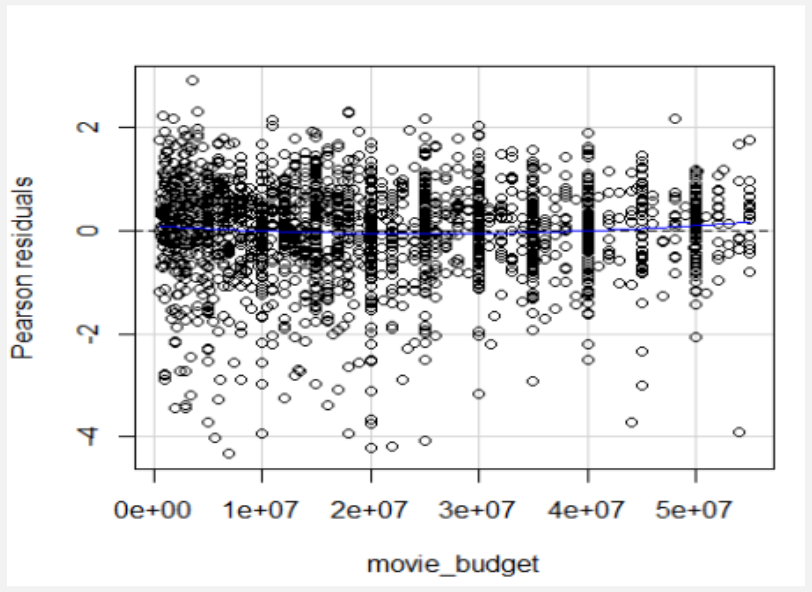
Appendix 5.2.: Heteroskedasticity Assessment with NCV Test Results – Final Model



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 114.526, Df = 1, p = < 2.22e-16
```

*This figure specifically examines the presence of heteroskedasticity within the context of the **final model** through the NCV test. The analysis is focused on different variables while the 'IMDb_score' serves as the dependent variable. The figure showcases the results of this test and underscores that it is conducted on the ultimate model, aiming to highlight the absence of any potential non-constant variance issue.*

Appendix 6: Residual Plot for Movie Budget Variable



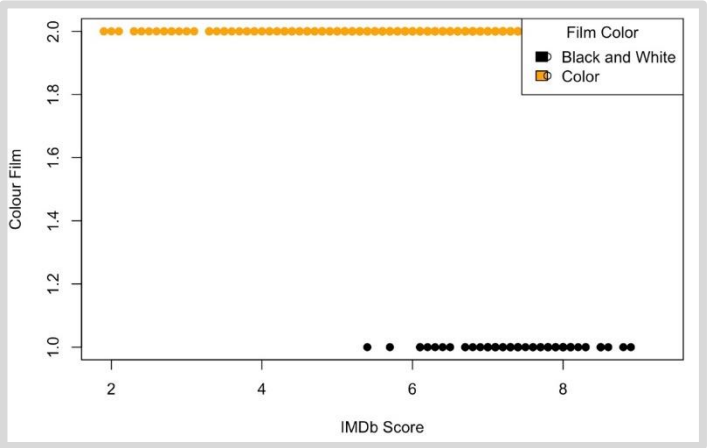
This figure presents a residual plot specifically focusing on the variable 'Movie Budget.' This figure presents a residual plot specifically focusing on the variable 'Movie Budget.'

Appendix 7: Predicted Scores from the Ultimate Regression Model

```
> predicted_scores
       1        2        3        4        5        6        7        8        9       10       11       12
5.086095 6.091991 4.124374 6.715702 5.956086 8.495828 7.742385 5.465918 5.964027 6.534381 6.681958 8.012047
```

This figure displays the forecasted scores generated by the final regression model, which has been implemented to predict the IMDb scores of the 12 movies under analysis. It provides a visual representation of the model's predictions for each of these movies, offering insights into its performance and accuracy in estimating IMDb scores.

Appendix 8 : Color vs. IMDb Score Scatter Plot



This figure is a scatter plot that visually represents the relationship between film color, categorized as a dummy variable (color or black & white), and IMDb scores. The Y-axis denotes the film color, and the X-axis represents the IMDb score for each film. This plot allows for an immediate visual assessment of how film color relates to IMDb ratings and provides insights into potential patterns or trends in the data.

Appendix 9: Regression Summary for Reg17

```
Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        6.167e+00  2.391e-01  25.793  < 2e-16 ***
bs(release_year, knots = c(quantiles), degree = 1)1 -1.242e+00  2.103e-01  -5.904 4.19e-09 ***
bs(release_year, knots = c(quantiles), degree = 1)2 -1.343e+00  1.846e-01  -7.277 4.97e-13 ***
bs(release_year, knots = c(quantiles), degree = 1)3 -1.292e+00  1.903e-01  -6.788 1.51e-11 ***
bs(release_year, knots = c(quantiles), degree = 1)4 -1.377e+00  1.899e-01  -7.252 5.95e-13 ***
bs(release_year, knots = c(quantiles), degree = 1)5 -1.773e+00  2.182e-01  -8.125 7.96e-16 ***
duration                                           8.222e-03  1.035e-03   7.947 3.25e-15 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)1  2.914e-01  1.268e-01   2.299 0.021628 *
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)2  4.367e-01  1.027e-01   4.251 2.23e-05 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)3  7.586e-01  1.095e-01   6.929 5.77e-12 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)4  1.129e+00  1.066e-01  10.597  < 2e-16 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)5  2.545e+00  3.039e-01   8.372  < 2e-16 ***
nb_news_articles                                   4.654e-05  1.118e-05   4.164 3.27e-05 ***
nb_faces                                          -3.976e-02  9.341e-03  -4.256 2.18e-05 ***
action                                            -3.028e-01  5.418e-02  -5.589 2.61e-08 ***
adventure                                         -1.988e-01  6.237e-02  -3.187 0.001462 **
thriller                                          -1.012e-01  4.883e-02  -2.073 0.038348 *
musical                                           -1.633e-01  7.479e-02  -2.184 0.029091 *
romance                                           -1.956e-01  4.685e-02  -4.176 3.10e-05 ***
sport                                              1.787e-01  9.097e-02   1.964 0.049620 *
horror                                            -4.623e-01  6.777e-02  -6.822 1.20e-11 ***
drama                                              5.175e-01  4.445e-02  11.643  < 2e-16 ***
war                                                2.206e-01  1.045e-01   2.111 0.034874 *
animation                                          7.396e-01  1.899e-01   3.895 0.000102 ***
crime                                              1.325e-01  5.224e-02   2.537 0.011269 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This figure displays the summary of the regression reg 17 showing the slopes of each predictor and the significance of each.

Appendix 10: Heteroskedasticity Coefficient Test

```
> coeftest(reg17, vcov=vcovHC(reg17, type="HC1"))

t test of coefficients:

                                                    Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                                        6.1666e+00  2.1884e-01 28.1786 < 2.2e-16 ***
bs(release_year, knots = c(quantiles), degree = 1)1 -1.2415e+00  1.7800e-01 -6.9749 4.212e-12 ***
bs(release_year, knots = c(quantiles), degree = 1)2 -1.3432e+00  1.5463e-01 -8.6865 < 2.2e-16 ***
bs(release_year, knots = c(quantiles), degree = 1)3 -1.2921e+00  1.5990e-01 -8.0807 1.134e-15 ***
bs(release_year, knots = c(quantiles), degree = 1)4 -1.3770e+00  1.5899e-01 -8.6608 < 2.2e-16 ***
bs(release_year, knots = c(quantiles), degree = 1)5 -1.7729e+00  1.9327e-01 -9.1734 < 2.2e-16 ***
duration                                           8.2219e-03  1.0658e-03  7.7143 1.951e-14 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)1  2.9141e-01  1.5504e-01  1.8796 0.0603181 .
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)2  4.3670e-01  1.2893e-01  3.3871 0.0007207 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)3  7.5858e-01  1.3322e-01  5.6942 1.433e-08 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)4  1.1293e+00  1.3349e-01  8.4594 < 2.2e-16 ***
bs(inv_movie_meter, knots = c(quantiles1), degree = 1)5  2.5446e+00  3.3721e-01  7.5458 6.931e-14 ***
nb_news_articles                                   4.6544e-05  1.8089e-05  2.5730 0.0101580 *
nb_faces                                          -3.9759e-02  1.0288e-02 -3.8646 0.0001150 ***
action                                            -3.0278e-01  5.8109e-02 -5.2106 2.087e-07 ***
adventure                                         -1.9876e-01  7.1676e-02 -2.7730 0.0056078 **
thriller                                          -1.0121e-01  4.8186e-02 -2.1005 0.0358201 *
musical                                           -1.6333e-01  8.1917e-02 -1.9938 0.0463151 *
romance                                           -1.9562e-01  4.5353e-02 -4.3132 1.692e-05 ***
sport                                              1.7871e-01  7.1672e-02  2.4935 0.0127343 *
horror                                            -4.6233e-01  7.3253e-02 -6.3114 3.433e-10 ***
drama                                              5.1749e-01  4.6506e-02 11.1274 < 2.2e-16 ***
war                                                2.2057e-01  9.2491e-02  2.3848 0.0171869 *
animation                                          7.3960e-01  1.7399e-01  4.2509 2.233e-05 ***
crime                                              1.3252e-01  5.2356e-02  2.5311 0.0114499 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This figure is showing the coefficient test to correct for heteroskedasticity

Appendix 11: Regression Summary for Reg12

```
Call:
glm(formula = imdb_score ~ bs(release_year, knots = c(quantiles),
    degree = 1) + duration + movie_budget + inv_movie_meter +
    nb_news_articles + nb_faces, data = imdb4)

Coefficients: (1 not defined because of singularities)
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        3.599e+00  2.077e-01  17.331  < 2e-16 ***
bs(release_year, knots = c(quantiles), degree = 1)1  3.648e-01  1.577e-01   2.313 0.020856 *
bs(release_year, knots = c(quantiles), degree = 1)2  2.483e-01  1.458e-01   1.702 0.088901 .
bs(release_year, knots = c(quantiles), degree = 1)3  2.873e-01  1.368e-01   2.101 0.035786 *
bs(release_year, knots = c(quantiles), degree = 1)4  1.416e-01  1.719e-01   0.824 0.410220
bs(release_year, knots = c(quantiles), degree = 1)5         NA         NA      NA       NA
duration                                           2.508e-02  1.477e-03  16.983  < 2e-16 ***
movie_budget                                      -1.235e-08  1.611e-09  -7.669 3.07e-14 ***
inv_movie_meter                                    4.985e+02  6.418e+01   7.767 1.46e-14 ***
nb_news_articles                                   1.346e-04  2.052e-05   6.558 7.44e-11 ***
nb_faces                                          -3.998e-02  1.075e-02  -3.720 0.000207 ***
```

This figure is showing the summary of reg12 where we were testing to see the effect of cutting of all the movies prior to 1995

Appendix 12: Pearson Residuals of Numerical Variables (2)

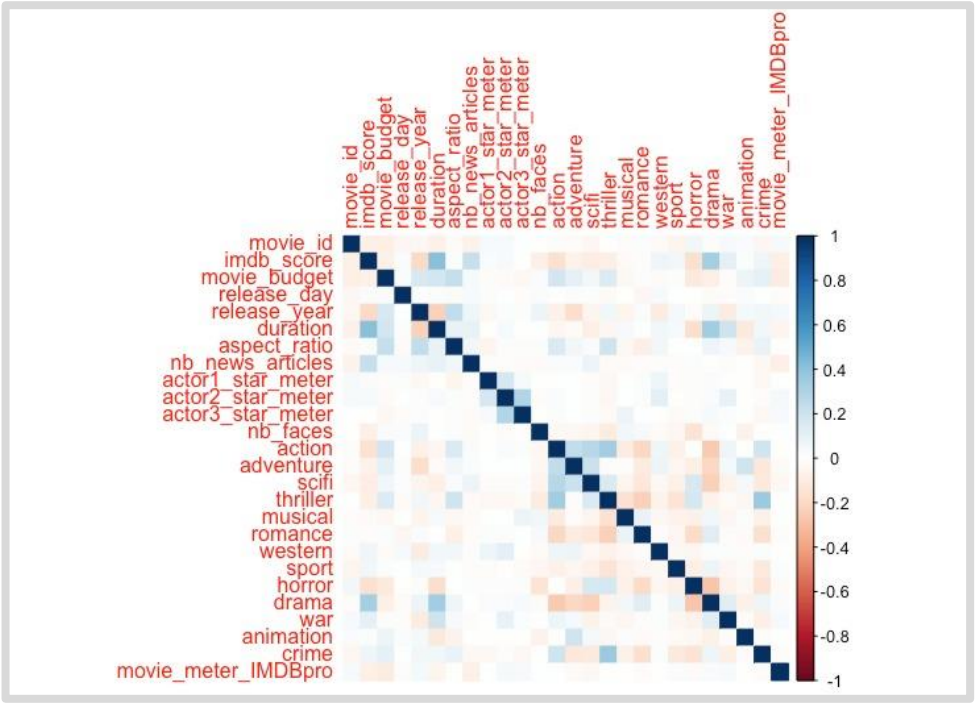
Pearson Residuals of Numerical Variables

	Dependent variable:
	IMDb Score
Movie budget	-0.000***
	(0.000)
Release year	-0.008***
	(0.002)
Duration	0.021***
	(0.001)
Aspect ratio	0.079
	(0.084)
Number of articles	0.0001***
	(0.00001)
Number of faces	-0.041***
	(0.011)
Movie meter IMDb pro	-0.00000***
	(0.00000)
Intercept	21.020***
	(4.016)
Observations	1,930
R^2	0.248
Adjusted R^2	0.245
Residual Std. Error	0.956 (df = 1922)
F Statistic	90.662*** (df = 7; 1922)
Note:	*p<0.1; **p<0.05; ***p<0.01

Pearson Residuals of Numerical Variables

	Dependent variable:
	IMDb Score
Release year	-0.019***
	(0.002)
Aspect ratio	0.248***
	(0.092)
Number of faces	-0.040***
	(0.012)
Intercept	44.202***
	(4.253)
Observations	1,930
R^2	0.047
Adjusted R^2	0.046
Residual Std. Error	1.075 (df = 1926)
F Statistic	31.784*** (df = 3; 1926)
Note:	*p<0.1; **p<0.05; ***p<0.01

Appendix 13: Complementary Correlation Matrix



Appendix 14: Supplemental Correlation Matrix with Distributional Results of Numerical Variables