# The 2021 IMDb Prediction Challenge

**Group 14: R Stars**

## Introduction:

It is often hard to predict what it takes to make a critically acclaimed movie. Do we need an Oscar-winning actor for it or should we be hiring multiple directors for it? Would adding multiple production companies in the mix to increase the budget make a difference? Or should we center the story in a way that it utilizes different languages so that it can cater to a wider audience?

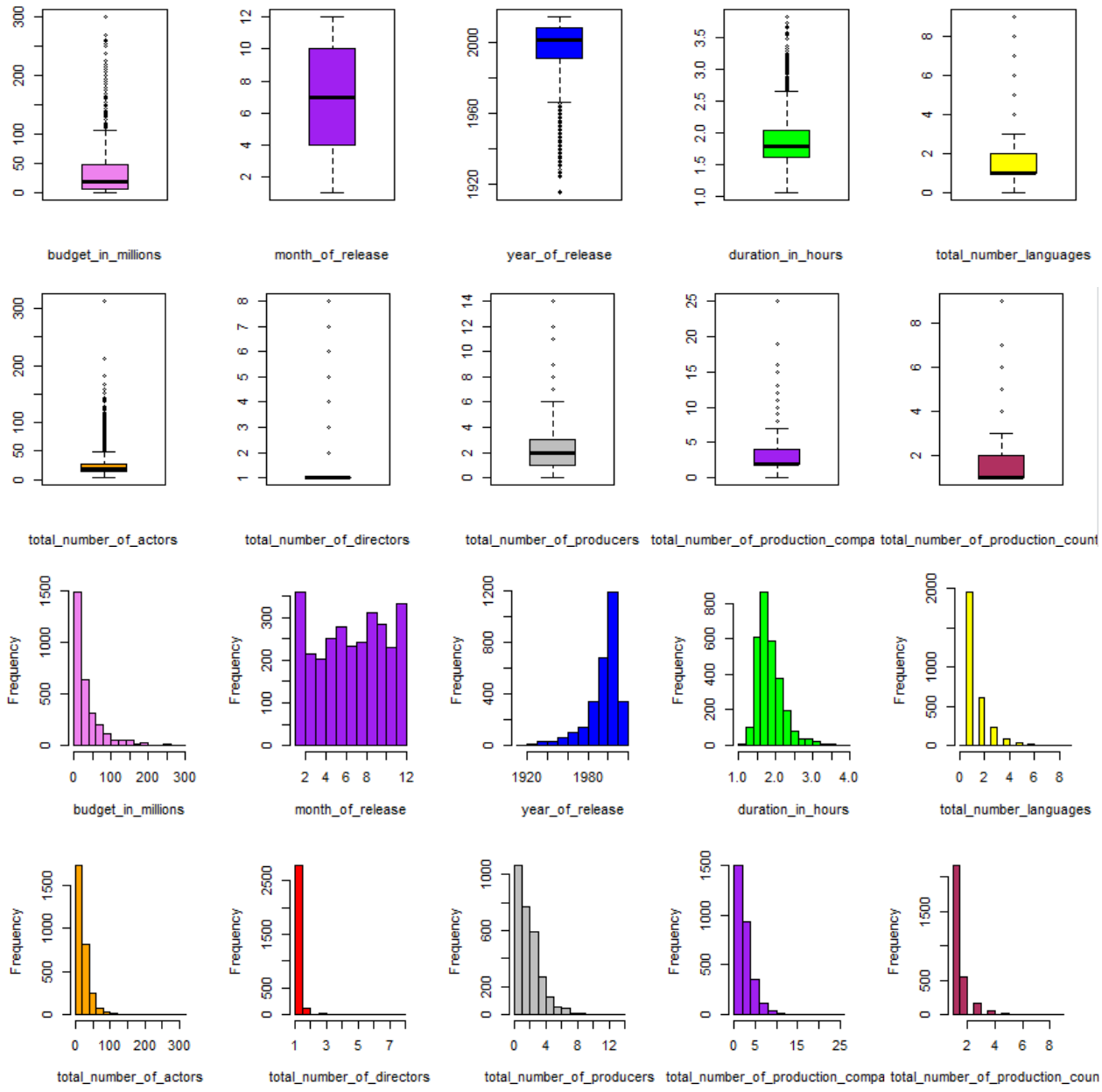
To this day, no one has been able to say for sure (with a confidence interval of 100%) what it really takes to put together a critically acclaimed movie, but using the power of predictive analytics we intend to make calculated estimates regarding how various factors can help us predict the critic scores or review scores for different movies.

Our main source of data was the IMDb database which was utilized to first evaluate each predictor and then find their degree of contribution to the prediction. Our target variable was imdb_score which is the average of the ratings given by each user on the website.

In the end, after evaluating each predictor and fixing any issues with collinearity, heteroskedasticity, and outliers we created a model while making sure that it wasn't under-fitted or overfitted according to our dataset.

## Data Description:

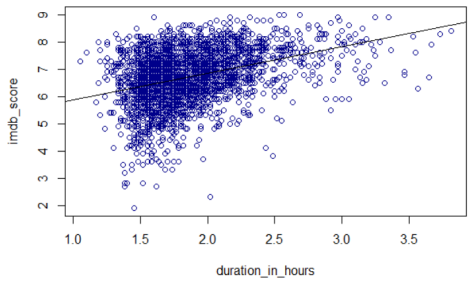
We performed exploratory analysis to understand the attributes of data. We visualized the distribution of our numeric predictors, followed by plotting their relationship with the target variable. For each predictor variable, we plotted the boxplot diagram to determine the range and outlier values. We later removed outliers during the model selection process. Histograms were also plotted in order to check the skewness and distribution of the dataset.




From the results, we noted that the budget, movie duration, number of languages, number of actors, number of directors, number of producers, number of production companies and number of production companies are right skewed. Year of release is left skewed, since the dataset contains more movies from recent years.

We ran simple linear regression between imdb_score and each of the numerical predictors. This was done to check for linearity trends and relationships between each predictor and imdb_score. This analysis was followed by running a multiple linear regression model performing a simple Tukey Test. We identified the predictors which showed a linear relationship

with the target variable i.e. imdb_score. Below is a sample linear regression graph between duration_in_hours and imdb_score, and the results obtained from the tukey test for all numerical variables. Please refer to *Appendix 1* for the simple linear regression graphs for all predictors.

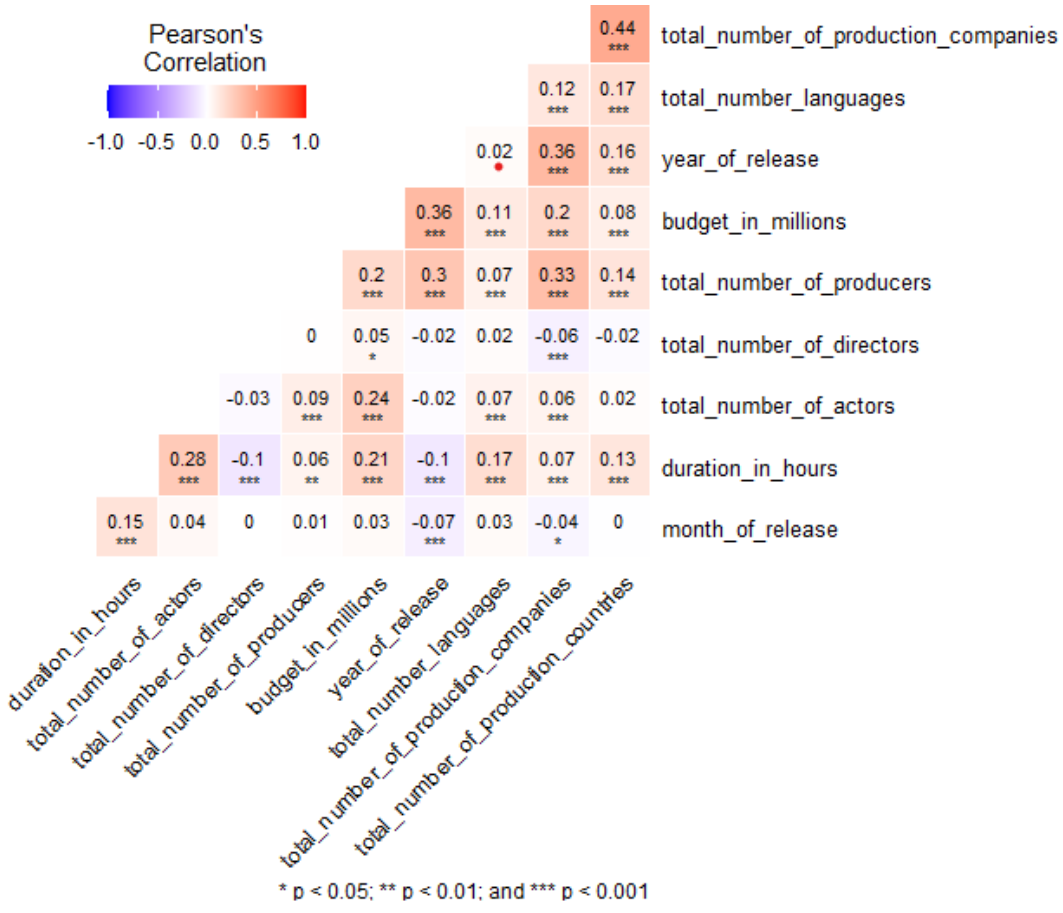



| | Test stat | Pr(>\|Test stat\|) | |
|---|---|---|---|
| budget_in_millions | 9.3250 | < 2.2e-16 | *** |
| month_of_release | -0.1343 | 0.893183 | |
| year_of_release | 4.5776 | 4.898e-06 | *** |
| duration_in_hours | -5.6355 | 1.911e-08 | *** |
| total_number_languages | -0.6748 | 0.499845 | |
| main_actor1_is_female | -0.7713 | 0.440612 | |
| main_actor2_is_female | 0.0723 | 0.942365 | |
| main_actor3_is_female | -1.9399 | 0.052492 | . |
| total_number_of_actors | -5.6690 | 1.576e-08 | *** |
| main_director_is_female | 1.3703 | 0.170707 | |
| total_number_of_directors | -2.2141 | 0.026899 | * |
| total_number_of_producers | -2.6813 | 0.007376 | ** |
| total_number_of_production_companies | -0.9041 | 0.366016 | |
| total_number_of_production_countries | -2.2832 | 0.022491 | * |
| Tukey test | -7.4588 | 8.729e-14 | *** |

---

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

A collinearity matrix of the entire numerical dataset was plotted in order to check for variables that are heavily correlated.



We noted that there were no pairs with an absolute value of correlation greater than 0.8. Therefore, we concluded that there is an insignificant correlation between the numerical variables, and we did not remove any variables based on the collinearity results.

# Model Selection:

After performing the exploratory analysis, we performed additional analyses to determine the best model for predicting the imdb_score. Our analysis included the below steps:

1. Detection of heteroskedastic predictors
2. Identification of interaction variable pairs
3. Determination of polynomial degrees for non-linear predictors using nested for loops and ANOVA test
4. Outlier analysis
5. Addition of relevant categorical predictors to the model

## 1. Detection of heteroskedastic predictors:

Before we could select the optimal modeling technique, it was important to refine our predictors and decide upon which predictors to include in our model. For this we ran simple linear regression models of each predictor in our dataset individually with the target variable and calculated heteroskedasticity using the ncvTest. We corrected heteroskedasticity from the variables and removed the variables where heteroskedasticity could not be corrected. This validated our assumption that the variance distribution is constant across the linear regression models. The following variables were removed as they were deemed to be insignificant: genre_animation, genre_sport, main_actor1_name, main_actor2_name, main_actor3_name, main_director_is_female, editor_name, total_number_of_production_companies, total_number_of_production_countries, genre_adventure, genre_music, genre_mystery, genre_romance, genre_thriller, main_director_name, and total_number_of_producers.

## 2. Identification of interaction variable pairs:

Next, we identified interaction terms to account for effects of variables on each other while predicting the imdb_score, for e.g. duration_in_hours and genre_drama. These interaction terms helped better explain the behaviour of variables with reference to the target variable, helping in increasing the model performance. This helps if one variable is not capturing the variance in the model. Combining it with other variables helps boost the R-squared value as it influences the effect of the variables on the model more. Please refer to *Appendix 2* for plots for 2 of the 9 identified interaction variable pairs.

## 3. Determination of polynomial degrees for non-linear predictors using nested for loops and ANOVA test:

We started building the model by running multiple linear regression, where we integrated all the relevant predictors. From the previous linear regression results and tukey test analyses, we concluded that a linear regression model would not serve as the best prediction model. We used a K-Fold test as our main cross-validation methodology, through which we calculated the model MSE. We noted that the MSE was not low enough, and hence decided to move to a multiple polynomial regression model.

For all the non-linear predictors identified in the Tukey test, we ran nested loops to find the optimal values for the degrees (between 1 to 5) of the non-linear predictors. We calculated the value of the MSE for each iteration using the K-Fold test, and noted the optimal degrees MSE of the model was the lowest. This led to the MSE dropping significantly causing the model performance to increase as compared to a multiple linear regression model. To further confirm our findings, we ran ANOVA tests on the regression models for different polynomial degrees to ensure that the chosen degrees are actually improving the model accuracy. Running a model with our finalized degrees of predictors, we achieved the best performing model with the lowest MSE. Finalizing these numeric predictors, we decided to fine-tune the model in order to bring down the MSE as much as we could, while ensuring that our model did not result in overfitting.

## 4. Outlier analysis:

We identified outliers using the Bonferroni outlier test and qqPlot. We removed the outliers iteratively and checked the MSE and Adjusted R-squared values at each step. We ensured that all the outliers removed were beyond 3*(standard deviation) to ascertain that the values removed are actually outliers. We removed 1-2% of the total data points, reducing the row count from 2953 to 2904. After every iteration, we evaluated the model's adjusted R-squared value and noted that the R-squared value increased significantly after removing the outliers.

Bonferroni outlier test provided us with observation numbers and their corresponding p-value of less than 0.05. Please refer to *Appendix 3* for the observation numbers and their p-values.

## 5. Addition of relevant categorical predictors to the model:

After finalizing the numeric predictors, we analyzed the categorical variables. We aggregated the data based on the frequency of each value. For a majority of the categorical variables, we noted that a large number of categories are occurring less than 10 times. Adding these predictors directly into the model might result in overfitting the model. Therefore, we categorized the less occurring categories into a single category called "Others". We tried to incorporate categorical variables and checked the model performance. The categories like main_actor1_name and main_director_name resulted in high values of MSE because of the large number of "Others" values. We finalized three categorical variables: main_production_company, main_production_country, main_lang. Along with running the K-Fold test, we also performed the Validation Set test to ensure that the model would work well with 40 data rows. Additionally, we added if-else conditions to check if the test set has a significant number of "Others" values for a particular categorical variable. In such scenarios, that categorical variable would be removed from the model automatically. Please refer to *Appendix 4* for the final predictors and their respective degrees.

Below are the results obtained from the final model tested on a random sample of 41 rows:

```
Residual standard error: 0.6643 on 2743 degrees of freedom
Multiple R-squared:  0.4797,    Adjusted R-squared:  0.4571
F-statistic: 21.25 on 119 and 2743 DF,  p-value: < 2.2e-16
```

```
> MSE
[1] 0.3722351
```

# Managerial Implications:

Given our model, we were able to highlight different attributes of a movie that help it get a higher imdb_score as compared to other movies. The main factors influencing the overall imdb_score of a movie would be budget_in_millions, year_of_release, duration_in_hours, total_number_of_languages, total_number_of_actors, total_number_of_directors, month_of_release along with a variety of genres and whether or not the main actor is a female or not.

Some variables were excluded in the analysis as they did not contribute to the prediction significantly, these included Reality TV genre, Short Film genre as they were single valued attributes so hence we could not conclude how they would influence the prediction.

If a producer was looking to make a critically acclaimed movie, one thing to keep in mind is that the overall budget allocated for the movie is very important in predicting its critical success as it influences a number of other factors as well. The range of the budget should be between 200 to 250 million USD in order to score the highest. Duration in hours of the movie is also an important factor to keep in mind while developing the screenplay. Movies with a runtime of between 2 to 3 hours fared the highest for imdb_score. The higher the number of languages integrated into the story of the movie improved the overall imdb_score as with the number of directors involved.

If the producer was picking a genre to produce a movie in, some genres fared better than others. Producing an action, biography, comedy, crime, documentary, drama, family, fantasy, film noir, history, horror, musical, scifi, war or eastern would help the movie perform better as compared to other genres.

An in-depth look at the interaction variables show us different correlations that contribute to better imdb_scores. For e.g. movies in the fantasy genre score better if they’re animated as compared to real-life movies, vice versa also being true. Thriller combined with drama genres fared better as compared to thrillers that weren’t dramas, with vice versa also being true. So all the interaction variables provide us with the genre combinations that help boost the imdb_score of a movie.

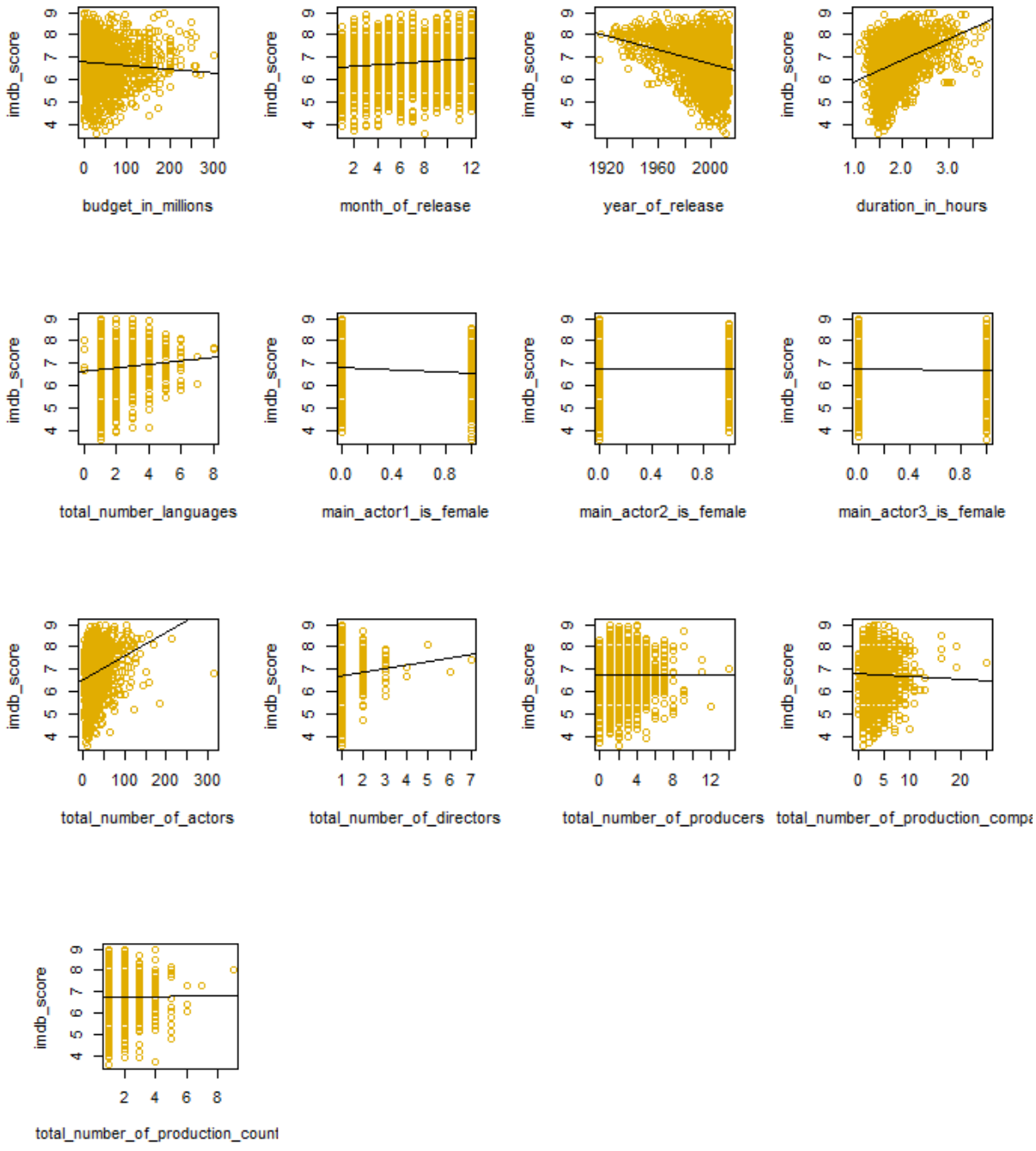
Having a female as the lead actor, may it be the first one, second one or the third one would definitely help increase the imdb_score, as a higher proportion of female led movies have historically scored better reviews and ratings as compared to male led movies.

As for the categorical variables, grouped according to their frequency, certain languages performed better than others when being rated, while some production houses had higher ratings as well, so for actors and directors it might be important to consider which production company they work with in order to get the highest critic scores.

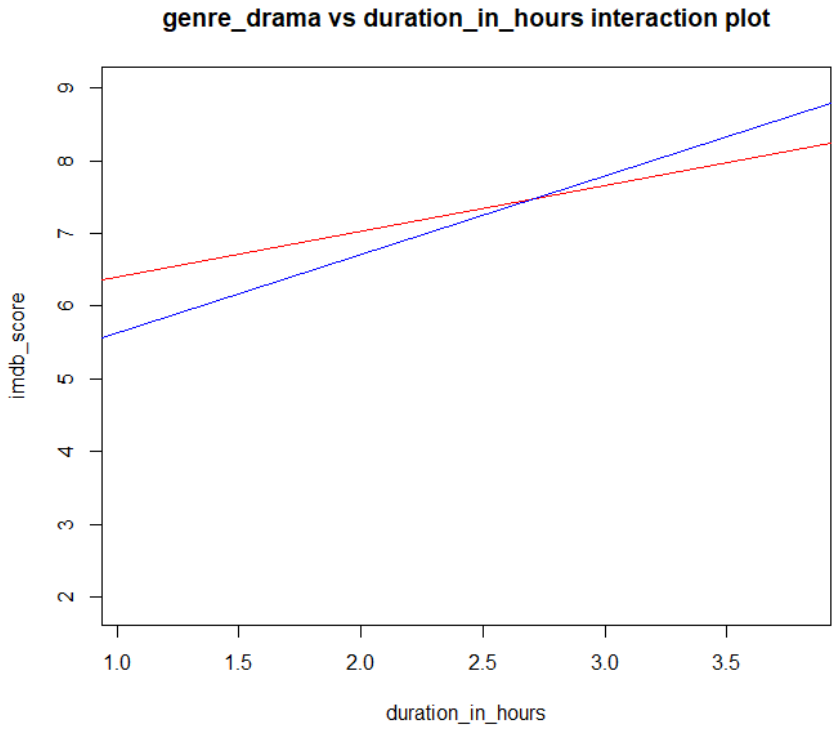
These insights may be able to help out actors struggling to catch a break, for e.g. Nicolas Cage, in opting for better movies, with a better chance of success. Or they may be able to help production companies set up better screenplays that fare well in terms of scoring.

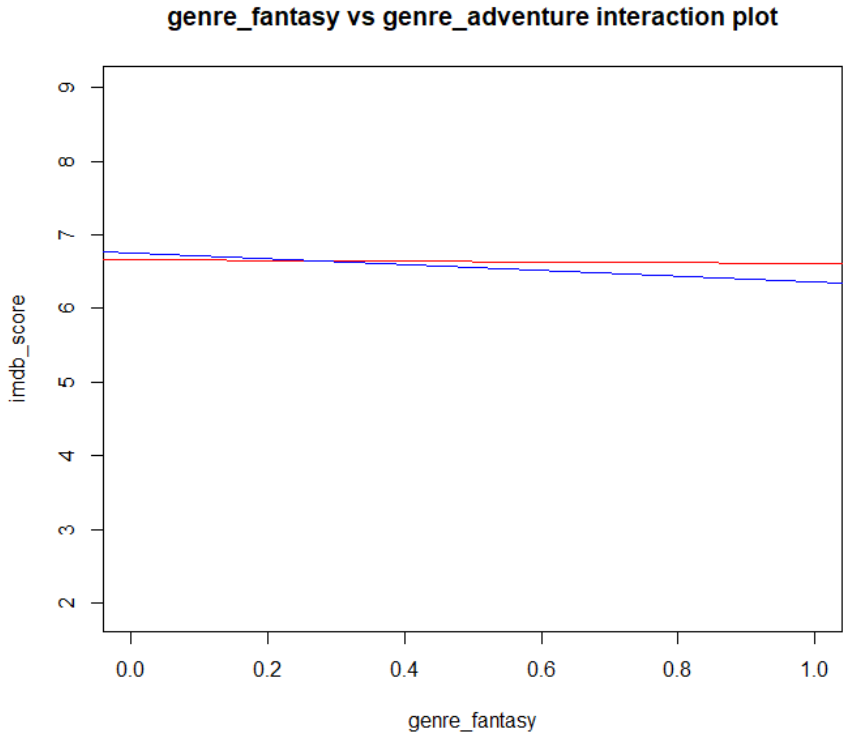
# APPENDIX:

### Appendix 1: Simple Linear Regression Plots



**Appendix 2: 2 of the 9 identified interaction variable pairs**



**Appendix 3: Outliers basis Bonferroni P-values**

| Observation Number | Bonferroni P-value |
|---|---|
| 633, 895, 2045 | P< 0.05 |
| 2718,2310 | |
| 2610,526 | |
| 574, 754 | |
| 446, 1162 | |
| 439, 1375 | |
| 264, 550 | |
| 2204, 2801 | |
| 431, 1904 | |
| 1889, 1969 | |
| 1143, 2044 | |
| 944, 1460 | |
| 898, 1968 | |
| 1498, 2654 | |
| 1445, 1542 | |
| 114, 751 | |
| 870, 1206 | |
| 1422, 2020 | |
| 8, 674 | |
| 1960, 2059 | |
| 343, 1888 | |
| 912, 1205 | |
| 258, 1921 | |

**Appendix 4: Final predictors and their degrees**

| Predictor | Degree |
|---|---|
| budget_in_millions | 4 |
| year_of_release | 3 |
| month_of_release | 3 |
| duration_in_hours | 4 |
| total_number_of_languages | 1 |
| total_number_of_directors | 1 |
| total_number_of_actors | 2 |
| total_number_of_production_companies | 2 |
| total_number_of_production_countries | 1 |
| total_number_of_production_countries | 1 |
| Genre: genre_action + genre_biography + genre_comedy + genre_crime + genre_documentary + genre_drama + genre_family + genre_fantasy + genre_filmnoir + genre_history + genre_horror + genre_musical + genre_scifi + genre_war + genre_western | 1 |
| main_actor1_is_female | 1 |
| main_actor2_is_female | 1 |
| main_actor3_is_female | 1 |
| Interaction Terms: (duration_in_hours*genre_drama) + (genre_drama*genre_romance) + (genre_drama*genre_thriller) + (genre_drama*genre_horror) + (genre_animation*genre_fantasy) + (genre_fantasy*genre_adventure) + (year_of_release*genre_drama) + (year_of_release*genre_comedy) + (year_of_release*genre_action) | 1 |
| Categorical variables: main_production_company_category + main_production_country_category + main_lang_category | 1 |