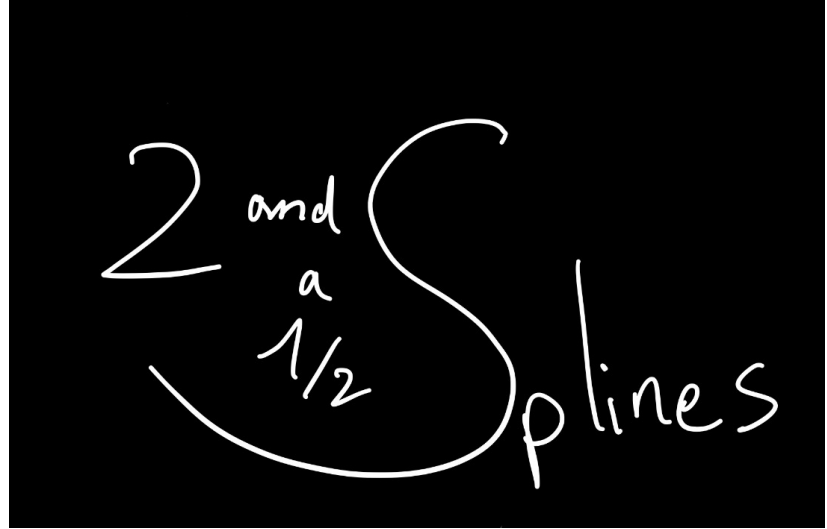


# Mid-term Project: The 2021 IMDB Prediction Challenge

## **Group 6: 2 and a ½ Splines**



Name: Matthieu Jacquand  
Student ID: 260728214

Name: Chelsea Hon  
Student ID: 261010089

Name: Fatima Nadeem  
Student ID: 260973683

Name: Hugo Garcia  
Student ID: 260791363

Name: Carlos Fabbri  
Student ID: 261018821

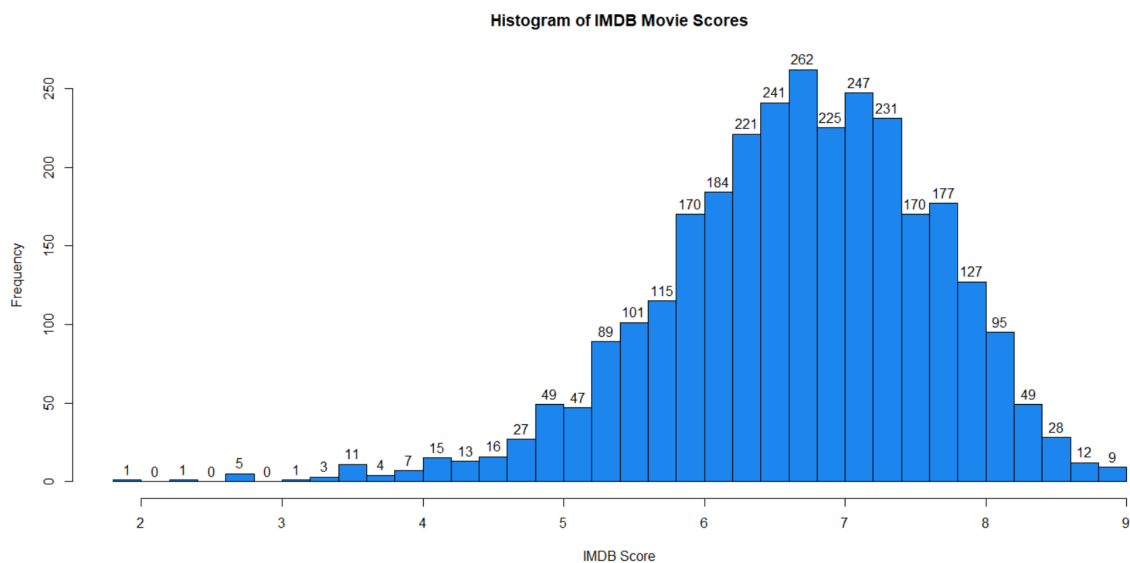
## 1. Introduction

Customer online reviews have opened up a new field in marketing and communications which transforms traditional word-of-mouth to a viral form of feedback that can influence consumer's opinions. On IMDb website (<https://www.imdb.com/>), the world's most popular and authoritative online source for movies, user rating simplifies the procedure in which users cast a vote (from 1 to 10) on every released movie title in the database. It is a great indicator for users to get inspired for the next movie, for actors to gain recognition standing in the eye of their fellow actors/ directors/ franchises, and for producers to build credibility, evaluate new movies and estimate the box office. In this project, we have built, trained, and deployed a statistical model with the objective to create the most powerful model to predict IMDb ratings of blockbuster movies and formulate the best recipe to create good movies. Ultimately, our goal is to predict the IMDb ratings of an unseen test dataset with good accuracy.

Using a given dataset of 2,953 movies from 1951 to 2014, as well as the tools learned in our class, we investigated characteristics of various movie attributes (also referred as predictors) by running tests to measure their performance power and then evaluating their overall worthiness for our model (see appendix 1 for data dictionary). Our group sought to approach this challenge using creativity, intuition, and a fusion of knowledge in statistics and business.

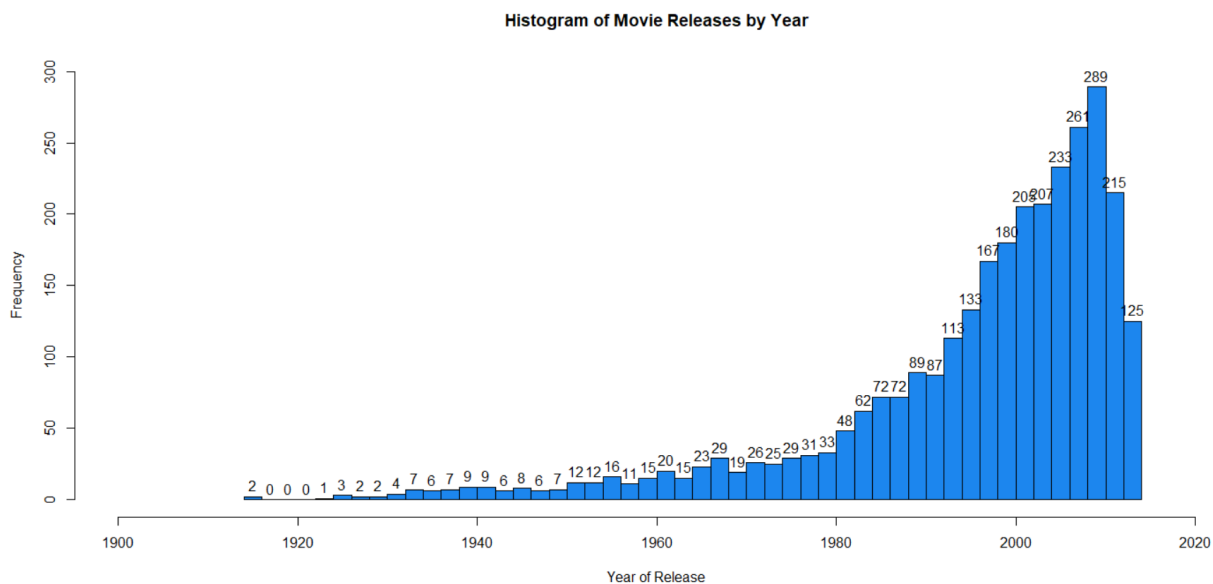
## 2. Data Description

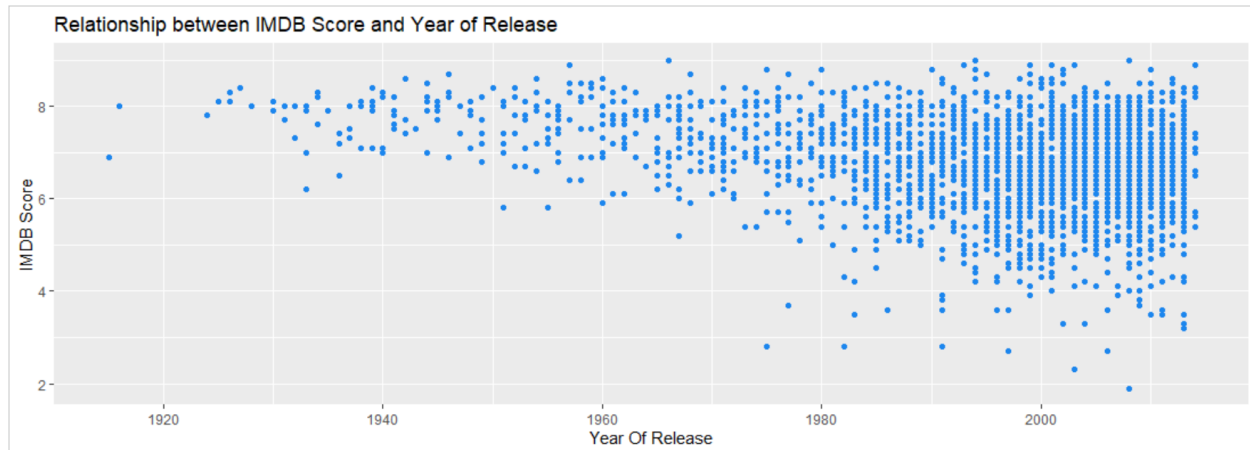
### Target Variable: IMDB Score



From the right-skewed normal distribution of IMDB score, we can construe that the majority of the movies in the dataset have been given a score between 5 and 8.

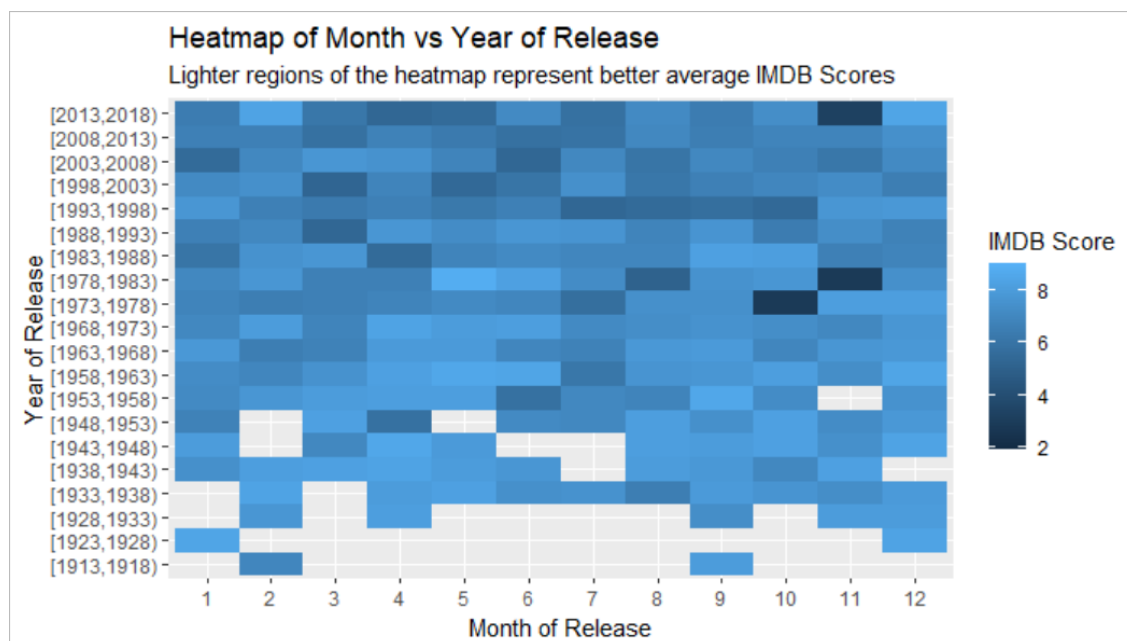
### Predictor: Year of Release





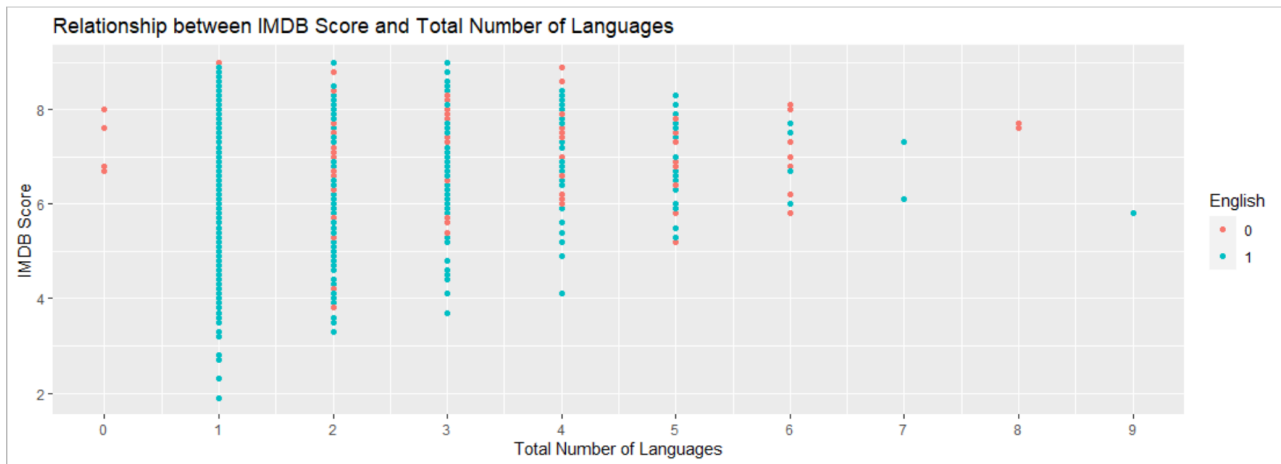
From the right-skewed distribution of the predictor Year of Release, it is observed that the majority of the movies present in the data set were released after 1980. The relationship between release year and IMDB score shown further portrays that the variance of IMDB scores increase over the years and movies released before 1950 have only received high IMDB scores. Hence, it can be concluded that a smaller sample of only positively perceived old movies are included in IMDB.

#### Predictor: Month of Release



The heat map above highlights that seasonality does not have a significant impact on the IMDB rating of a movie. Hence, we can infer that the month in which a movie is released does not contribute towards its IMDB rating. However, it should be noted that seasonality does have an impact on the sales and revenue generated for a movie which has not been examined by the model.

**Predictor: Total Number of Languages**

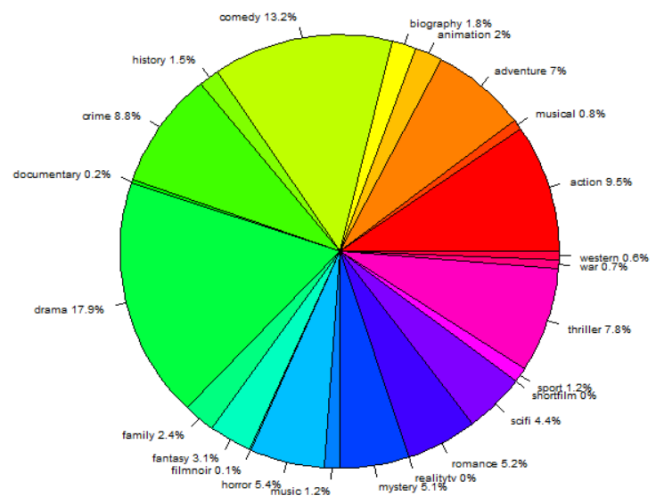


**See appendix 2 for Distribution for all Languages**

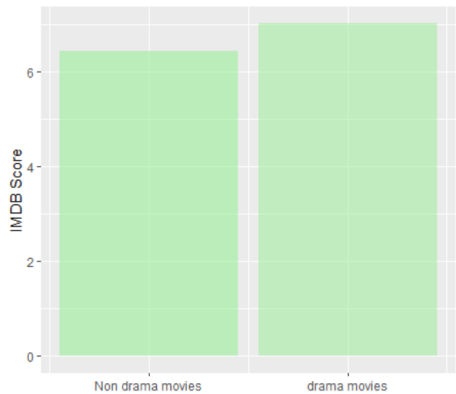
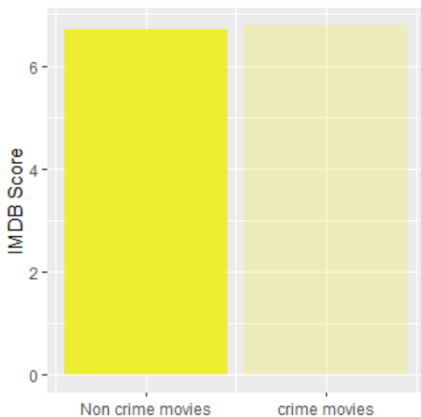
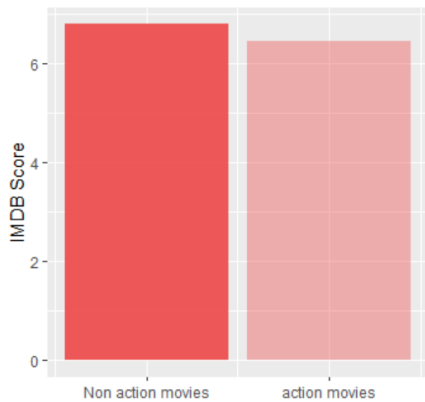
As English was the most dominant language, we used it as a binary variable (English = 1, All other Languages = 0) in our exploration of the relationship between IMDB Score and Total Number of Languages. As per our analysis, there is generally more variance in IMDB scores for movies whose main language is English. This allows us to infer that if a movie’s main language is not English and has been translated into more than one language, it is a good movie having a comparatively higher IMDB score.

**Predictor: Movie Genres**








Through analyzing the distribution of each type of genre in our data we were able to determine the most popular genres. A further exploration was conducted to examine the relationship between each popular genre with IMDB score alongside all other genres. Three distinct types were outlined and Drama, Crime and Action are portrayed as examples below to show each type.



**See appendix 3 for Genre Distribution chart**

Drama	Crime	Action																		
Movies of genre Drama tend to have higher IMDB Scores than the average	Movies of genre Crime tend to have a slightly higher Score than the average.	Movies of genre Action tend to have lower IMDB Scores than average																		
 <table><tr><th>Category</th><th>IMDB Score</th></tr><tr><td>Non drama movies</td><td>6.5</td></tr><tr><td>drama movies</td><td>7.0</td></tr></table>	Category	IMDB Score	Non drama movies	6.5	drama movies	7.0	 <table><tr><th>Category</th><th>IMDB Score</th></tr><tr><td>Non crime movies</td><td>6.8</td></tr><tr><td>crime movies</td><td>7.0</td></tr></table>	Category	IMDB Score	Non crime movies	6.8	crime movies	7.0	 <table><tr><th>Category</th><th>IMDB Score</th></tr><tr><td>Non action movies</td><td>6.8</td></tr><tr><td>action movies</td><td>6.5</td></tr></table>	Category	IMDB Score	Non action movies	6.8	action movies	6.5
Category	IMDB Score																			
Non drama movies	6.5																			
drama movies	7.0																			
Category	IMDB Score																			
Non crime movies	6.8																			
crime movies	7.0																			
Category	IMDB Score																			
Non action movies	6.8																			
action movies	6.5																			

### 3. Model Selection

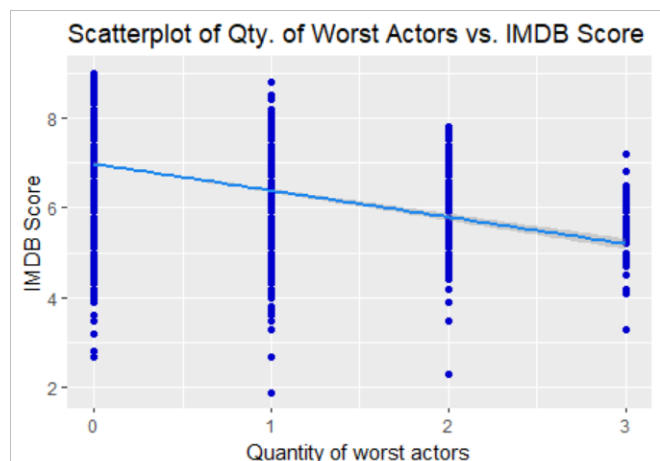
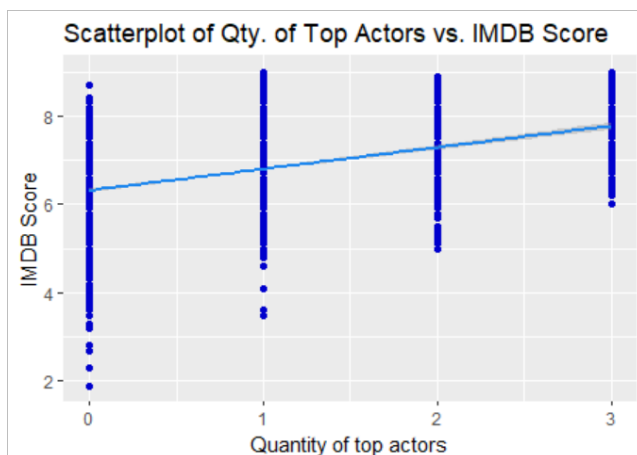
		# of Predictors	Approach	Model	Results	Remarks
	0. Data Exploration	50	Familiarizing and visualizing the dataset	N/A	N/A	N/A
	1. Feature Engineering	53	Deriving more predictive feature from top/worst actors/directors	N/A	N/A	best/worst actor/director (4)
	2. Data Preprocessing	131	Dummifying variables and enhancing data usability	N/A	N/A	Month of release (12-1), main lang (38-1), main production country (47-1)
	3. Model Base-lining	62	Adding up predictors to maximize adjusted r-squared	Multiple linear regression, with assumed linearity	R2 = 0.5161	budget year duration (3), genre (15), total num (5), best/worst actor/director (4), actor gender (3), month (5), lang (11), prod country (16)
	4. Testing	62	Evaluating linearity, Heteroskedasticity, removing outlier, and cross validation	Multiple linear regression, with outliers removed	R2 = 0.5299 MSE = 0.4382 MAPE = 7.961%	Non-linearity relationship to be explored
	5. Fine-tuning with Polynomial	62	Testing predictors for degree 1 to 3 to reduce MSE	Multiple poly regression, in optimal degree	4 poly terms MSE = 0.4124	duration (2), # of directors (3) & producers (3), qty of top actors (2)
	6. Fine-tuning with Splines	62	Plotting predictors distribution and test out spline knots to reduce MSE	Multiple poly regression with splines	3 splines MSE = 0.4075 MAPE = 7.652%	budget (.33, .66), # of actors (.1,.3,.5), # of prod comp (.33, .66)

*See appendix 4 for remarks elaboration*

#### 3.1 Feature Engineering

Before going into the actual model we focused our efforts on feature engineering and came up with new features derived from the dataset that were more predictive than any “raw” feature by itself.

Instead of using distinct 3,300+ actors (from main actor 1, 2, 3 fields) and 1,300+ directors dummified variables directly in a regression model, we extract a list of “top” and “worst actors”. For top actors, we consider the top 350 actors based on average IMDB score. Similarly, for worst actors, we consider the 240 least ranked according to average IMDB score. Such parameters are carefully selected in a way to optimize a maximum adjusted r-squared and a minimum MSE. Then, based on these lists, for each movie, we counted the quantity of top and worst actors that they include. A similar method is applied for directors and producers.



### 3.2 Data Preprocessing

Apart from this, we studied the data and performed a few steps to enhance the quality and usability of data fitting into the model. First, we dropped the categorical columns which either are too distinct, did not contain values, or have been addressed in feature engineering. Then, we dummified the categories variables `month_of_release`, `main_lang` and `main_production_country` by removing the first dummy. Since the R stat package does not take special characters and spacing in modelling, additional data cleaning on `main_lang` is needed before we can use them as predictors. With that, we were ready to model.

### 3.3 Model Base-lining & Selecting Predictors

We began first with a preliminary model with the goal of having a baseline. With a full list of 131 predictors, we built a loop to (i) understand the variance each single predictor can introduce to a simple linear regression against the `imdb_score` (the higher the r-squared the better) and (ii) determine whether adding that new predictor to a multiple linear regression can improve how well data points can fit into the line (i.e. if adjusted r-squared increases, keep the predictor, else drop it).

For example, the multiple linear regression,  $\text{imdb\_score} = b_0 + b_1 * \text{budget\_in\_millions} + b_2 * \text{year\_of\_release}$ , has an adjusted r-squared 0.077. Adding the third predictor will update the regression to,  $\text{imdb\_score} = b_0 + b_1 * \text{budget\_in\_millions} + \dots + b_3 * \text{duration\_in\_hours}$ , and increase the adjusted r-squared to 0.192. Hence, we can keep the `year_of_release` term. The similar goes until the 6th predictor, where  $\text{imdb\_score} = b_0 + b_1 * \text{budget\_in\_millions} + \dots + b_6 * \text{genre\_adventure}$ , decreases the adjusted r-squared from 0.21148 to 0.21125. We drop the term and move on.

After running 131 iterations, 62 predictors remained and the highest adjusted r-squared 0.516 was obtained. This served as our model baseline.



				MregAdj		
	Predictor	R_Squared	P_Value	Rsquares	MregPvalues	Decision
1	budget_in_millions	0.0044	0	0.0041	0	keep
2	year_of_release	0.0765	0.0003	0.0770	0.0003	keep
3	duration_in_hours	0.1307	9.14E-75	0.1920	3.70E-71	keep
4	total_number_languages	0.0067	6.35E-53	0.1929	0.0431	keep
5	genre_action	0.0228	0	0.2115	5.86E-51	keep
6	genre_adventure	0.0007	7.32E-92	0.2112	1.42E-45	drop
7	genre_animation	0.0004	0	0.2275	0.0007	keep
8	genre_biography	0.0217	9.06E-06	0.2314	2.83E-33	keep
...						
	main_production_country_unite					
131	d_states_of_america	0.0049	0	0.5161	0.007997	drop

### 3.4 Testing

Minimizing variance (Maximizing r-square) is not good enough as having too many terms in our model will run into the problem of over-fitting. A misleadingly high r-squared value can lead to misleading projections. Therefore, a series of testing and fine-tuning is needed.

Linearity test - the baseline model was based on the assumption that each predictor contributes to IMDB score to the power of one. To validate, we performed a residual plots analysis and observed curved patterns of residuals of 4 predictors denoting that linearity was being violated: budget\_in\_millions, duration\_in\_hours, total\_number\_of\_actors, total\_number\_of\_producers. Having p-value  $\sim 0$  (below 0.05) meant that the model did not pass the linearity test.

				MregAdj	Mreg	Pr(> Test	
	Predictor	R_Squared	P_Value	Rsquares	Pvalues	Decision	Test stat stat  )
1	budget_in_millions	0.0044	0	0.0041	0	keep	5.8994 4.07E-09
3	duration_in_hours	0.1307	9.14E-75	0.1920	3.70E-71	keep	-3.8397 0.0001
30	total_number_of_actors	0.0459	1.80E-07	0.3314	5.55E-34	keep	-4.8104 1.58E-06
33	total_number_of_producers	7.07E-06	0	0.3336	9.99E-14	keep	-4.8104 6.35E-06
	Tukey test						-5.2742 1.33E-07

Even when we removed the least linear variables, the overall p-value only increased to  $\sim 0.00076$ , which was not ideal. Therefore, we decided to keep the 62 predictors and fine tune the model with polynomials in the next section.

Heteroscedasticity Test - running `ncvTest`, we obtained a p-value of  $2.22e-16$  ( $<0.05$ ) and detected heteroscedasticity in our model where variance of errors is not constant and the linear regression will give biased standard errors. We attempted to correct this by transforming the regression standard errors but no significant improvement was found in the p-value. This non-constant variance gave us further evidence to explore non-linearity relationships between the predictors and imdb score in the next section.

Removing Outliers - applying a `qqPlot` visualization and Bonferroni test, 8 data points are identified as the outliers which have low probability of occurrence but have much influence to the regression line. The adjusted r-squared before and after were 0.5161 and 0.5299 respectively. This showed that our re-trained model is slightly more robust once we removed the outliers from the dataset.

Cross Validations - 10-fold cross validation was performed to assess the effectiveness of our model (the lower MSE the better). Data was divided randomly into 10 groups, in which 9 folds are training sets and 1 fold is testing alternating in each case. Followed by train-test split cross validation of 70% training and 30% testing for 10 times, we were able to measure the prediction accuracy of our model on imdb score on average.

The model performed fairly well with an MSE of 0.4382 and a MAPE (mean absolute percentage error) of  $\sim 7.961\%$ . This signified that our predictions are off by only  $\sim 8\%$  against the true values on average.

### 3.5 Fine tuning with polynomials

Having found the most relevant predictors for the `imdb_score`, we set out to improve our model accuracy by testing out different polynomial regressions for our assumed linear predictors. At this stage, binary predictors (i.e. dummies relevant to movie genre, language and production company), cannot be used in polynomial regression as they can only take two values (0 and 1). The list of predictors that can potentially be fine-tuned using polynomials are the following: budget, release year, duration and total number of languages, actors, directors, producers and production companies. This also included our newly engineered features: quantity of best and worst actors. Other predictors are still assumed as linear in the model.

To do so, we set up a loop to test all possible combinations of degrees for these 10 identified features alongside single degree linearity for the remaining 52 predictors. The main obstacle here was the computational power. We first set out to try degrees 1 to 4 for the first 8 polynomial predictors and 1 to 3 (as it is the maximum degree possible) for the final two polynomial predictors (quantity of top and worst actors). However, that meant trying out 589 824 ( $4$  to the power of  $8$  times  $3$  to the power of  $2$ ) models. Even by using 5-fold cross validation to compute

each tested model's mean squared error, enabling the loop to test each model in ~0.5 second, still displayed 3.5 days of runtime. Therefore, we decided to test all 10 polynomial predictors for degrees 1 to 3; that is testing 59,049 models in total. This loop ran in precisely 26,712 seconds; roughly 7 and a half hours.

From this loop, we were eventually able to extract the best degree parameters for the 10 tested polynomial predictors. However, it came to our notice that 7 of them work better with either quadratic or cubic degrees.

	CombinationMSE	DegreeCombination
42025	0.4124143	3 1 2 1 2 3 3 2 2 1
42016	0.4124381	3 1 2 1 2 3 3 1 2 1
42022	0.4125086	3 1 2 1 2 3 3 2 1 1
42013	0.4125373	3 1 2 1 2 3 3 1 1 1
42754	0.4125600	3 1 2 2 2 3 3 2 2 1
42745	0.4125749	3 1 2 2 2 3 3 1 2 1
42751	0.4126585	3 1 2 2 2 3 3 2 1 1
42742	0.4126786	3 1 2 2 2 3 3 1 1 1
48577	0.4126828	3 2 2 1 2 3 3 1 2 1
43483	0.4126936	3 1 2 3 2 3 3 2 2 1
48586	0.4127008	3 2 2 1 2 3 3 2 2 1
43474	0.4127110	3 1 2 3 2 3 3 1 2 1

```

> MSE_POLY_results3= do.call(rbind, Map(
>
>
> proc.time() - ptm
      user  system elapsed
26284.57   155.59  26712.30

```

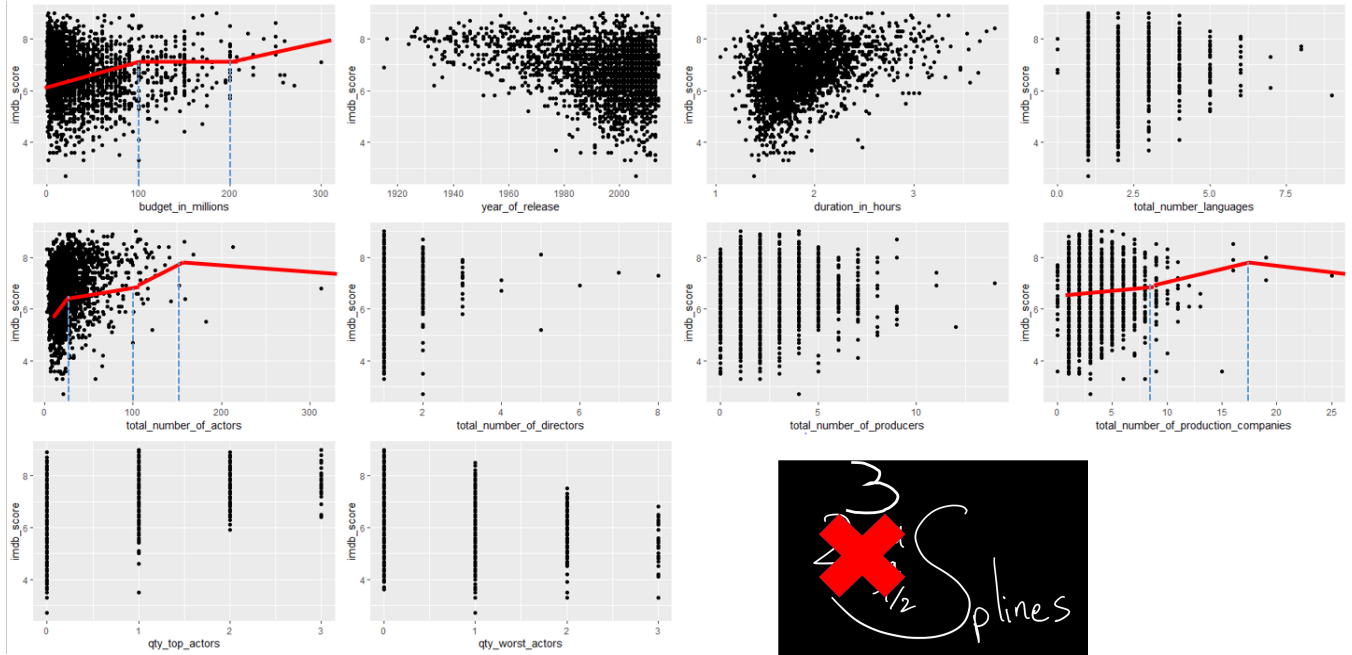
### 3.6 Fine tuning with splines

After further improving our MSE by 0.02 using polynomial regression, we decided to look into splines to further chase the best possible MSE. We began by plotting the distribution of these 10 predictors in relation to the imdb\_score.

From these plots, it wasn't at first evident that splines had to/could be used to improve model accuracy. The predictor for which it seemed most suitable was the total\_number\_of\_actors predictor that held the majority of its distribution for low values (not everyone can afford super production with 100+ actors). For this predictor, we decided to set knots at the .10, .30 and .50 quantiles.

For other predictors, it was a matter of trial and error to determine if adding spline knots at "traditional" quantiles (.33 & .66, .25 & .50 & .75) and trying out different combinations of splined and non-splined predictors could further reduce model MSE. In the end, the best model accuracy was obtained using splines for budget, number of actors and number of production companies' predictors, or 3 splines total.

*(We cannot help but observe that this is oddly close to our group name, maybe fate gave us a hint?)*



Finally, we run a final loop to test 125 models, testing different polynomial degree combinations (1 to 5) for the newly splined predictors, keeping previously determined degrees for other predictors. We managed to further reduce our MSE to 0.4075.

### 3.7 Final model results

We achieved the lowest MSE at: 0.4075 (based on 5-Fold cross validation, 0.4159 based on 10 fold cv). We used a total of 62 predictors, 55 of which are linear, 4 of which are polynomial with varying degrees and 3 of which are polynomial with varying degrees and varying spline knots. Please find the final formula for our model in appendix 5. Running a sample train\_test\_split with a 30% test size, we obtain a MAPE (Mean Absolute Percentage Error) of 7.652%.

With this same test set of 885 predictions, we further analyzed the performance of the model and attempted to understand which factors helped the model to make accurate guesses. For this, we grouped our predictions into quintiles according to the absolute difference between our predicted score and the real IMDB score. Based on these 5 quintiles, we could notice, for example, that our model tends to be accurate when predicting the scores of films directed by females (36% of these types of films belong to the top 20% best predictions). On the contrary, movies of the horror genre tend to make it harder for the model to produce accurate predictions (only 26% of these movies belong to the top 40% best predictions). Please refer to the table below to review other features that seem to rank the model's predictive power.

Quintiles according to accuracy of prediction	Delta (Predicted vs True IMDB Score)	Distribution of Films directed by females	Average of quantity of top actors	Average Number of Production Companies	Distribution of Horror Movies
1 (20 % Best predictions)	0.08	36.67%	0.46	3.10	13%
2	0.22	16.67%	0.46	3.02	13%
3	0.39	23.33%	0.42	3.01	19%
4	0.59	13.33%	0.42	2.95	28%
5 (20% Worst predictions)	1.06	10.00%	0.34	2.72	27%
Total	0.47	100%	0.42	2.96	100%

## 4. Managerial Implications

### **Budget Implications**

While there is a significant relationship between budget (in millions) and IMDB rating, it is not a linear relationship. So, spending more on a movie will not translate into a higher IMDB score. Of course, having a comfortable budget is still important as it can help secure reputable actors, pay for marketing and pre-release promotions, and pay for other talents such as editors and other directors, all of which may influence the IMDB score.

### **Main Actor is Female Implications**

As per our model findings, a significant negative relationship was found between female main actors and IMDB scores (See Appendix 9 for statistical supplement). Such results highlight the idea of a gender bias present in the entertainment industry which may prompt critics/users to award lower scores to movies with female protagonists. Nevertheless, it should be noted that our model is limited in this prediction as only movies released by 2014 are included in the analysis, and fewer movies with female actors were released by 2014. Currently, there is an increasing trend of female actors in the industry (Statista, 2021), and from a managerial perspective, the hidden potential in this area should be explored further.

### **Genre Implications**

In our model, movies in the animation, documentary, and drama genres are going to be rewarded with a higher prediction for IMDB score. Contrastingly, movies in the action, comedy, family, fantasy, and horror genres are going to be penalized with a lower prediction for IMDB score (See appendix 6 for statistical supplement).

**Managerial Implication:** Given the drama genre is popular in our dataset, occurring in nearly a fifth of movies in the dataset, and that it is favored by our prediction model, we would recommend pursuing this genre for a movie to maximize chances of earning a higher IMDB score.

### **Main production Country Implications:**

From our results, it can be inferred that depending on the country, the main production country of a movie can have an impact on a movie's IMDB rating (See appendix 7 for statistical supplement). Consequently, from a managerial perspective, countries which have significant positive effects on scores should be explored first as primary financial contributors (Iceland, Belgium, New Zealand, Germany, Philippines or the United Kingdom). Whereas Bulgaria, Ecuador, Norway and Switzerland have all been proven by our model to be the least favourable production countries to work with in terms of movie scores. However, it should be noted that our model explores only the effect on IMDB scores and not capital which is an important factor in the evaluation of a production country.

**Main Language Implications:**

From our model, we observe that movies with English as their main language will negatively impact the predicted IMDB score (See appendix 8 for statistical supplement). At first this observation seems counterintuitive, but since most movies in the dataset are in English (see Appendix 2 for distribution of languages in movies), their IMDB scores - both high and low rated - are being accounted for in our model. *Essentially:* Because movies with a main language not in English are in small majority, the ones that do make it onto IMDB's website, are more likely to be higher rated by our model. Managerial Implication: movies with their main language in English will get slightly penalized by our model. Should a director seek to maximize their chances of earning a high IMDB score with our model, we recommend they choose their main language to be in 'Nederland' or 'Suomi'.

5. Appendices

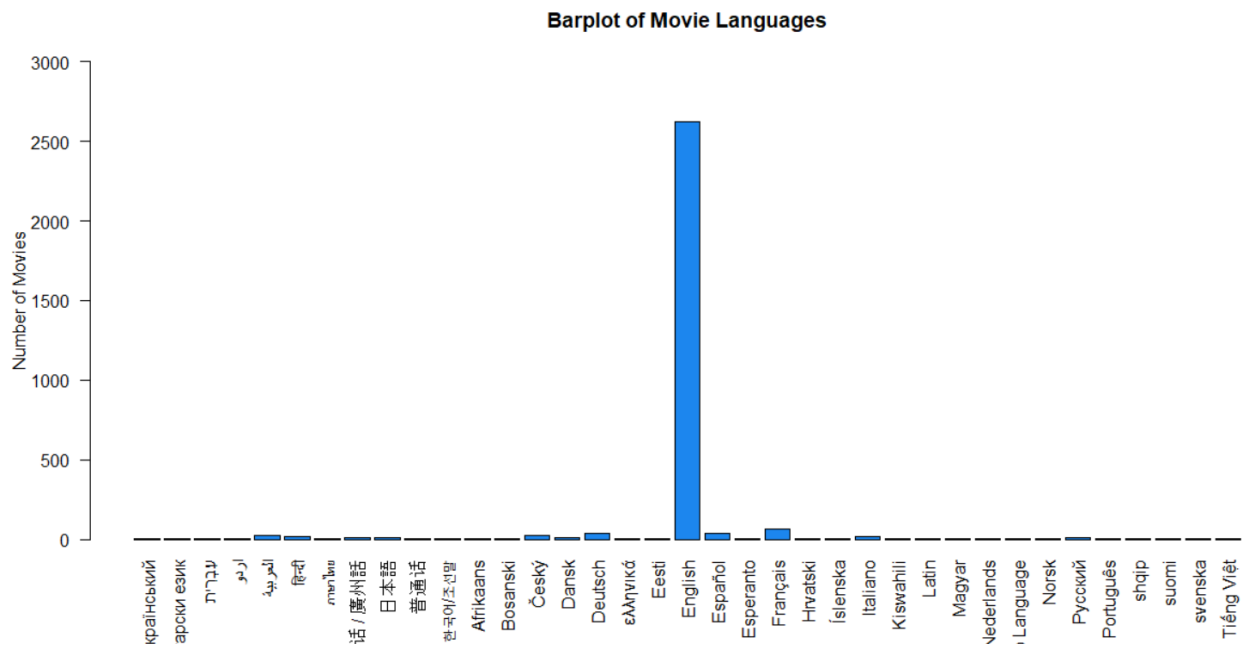
Appendix 1 - Data Dictionary

Labels	Dependent Variable
imdb_id	imdb_score
imdb_url	
title	

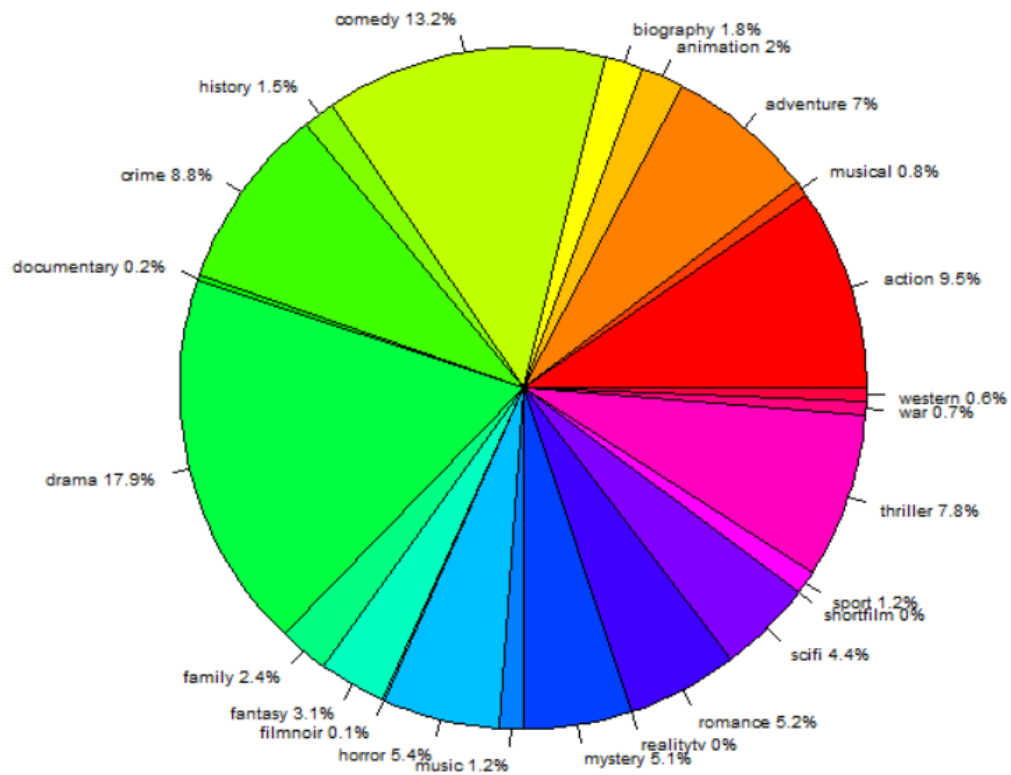
Predictors

budget_in_millions (of US dollars)	genre_filmnoir (0 or 1)	main_actor3_name
month_of_release	genre_history (0 or 1)	main_actor1_is_female (0 or 1)
year_of_release	genre_horror (0 or 1)	main_actor2_is_female (0 or 1)
duration_in_hours	genre_musical (0 or 1)	main_actor3_is_female (0 or 1)
main_lang (main language of the movie)	genre_mystery (0 or 1)	total_number_of_actors
total_number_languages	genre_realitytv (0 or 1)	main_director_name
gente_action (0 or 1)	genre_romance (0 or 1)	main_director_is_female
genre_adventure (0 or 1)	genre_scifi (0 or 1)	total_number_of_directors
genre_animation (0 or 1)	fenre_shortfilm (0 or 1)	main_producer_name
genre_biography (0 or 1)	genre_sport (0 or 1)	total_number_of_producers
genre_comedy (0 or 1)	genre_thriller (0 or 1)	editor_name
genre_cime (0 or 1)	genre_war (0 or 1)	main_production_company
genre_documentary (0 or 1)	genre_western (0 or 1)	total_number_of_production_companies
genre_drama (0 or 1)	main_actor1_name	main_production_country
genre_family (0 or 1)	main_actor2_name	Total_number_of_production_countries
genre_fantasy (0 or 1)		

Appendix 2 - Bar plot showing distribution of the main languages in movies



### Appendix 3 - Pie Chart of Genre Distributions



### Appendix 4 - Modelling Overview Remarks Elaboration

Stage	# of Predictors	Remarks	Elaboration for “( )”
1. Feature Engineering	53	best/worst actor/director (4)	Number of related predictors
2. Data Preprocessing	131	Month of release (12-1), main lang (38-1), main production country (47-1)	Number of dummified predictors removing the first one
3. Model Base-lining	62	budget year duration (3), genre (15), total num (5), best/worst actor/director (4), actor gender (3), month (5), lang (11), prod country (16)	Number of related predictors



5. Fine-tuning with Polynomial	62	duration (2), # of directors (3) & producers (3), qty of top actors (2)	Degree of polynomial
6. Fine-tuning with Splines	62	budget (.33, .66), # of actors (.1,.3,.5), # of prod comp (.33, .66)	Decile for spline knots

#### Appendix 5 - Best achieved model formula

```
fbest = as.formula(imdb_score ~ bs(budget_in_millions, knots = c(quantile(budget_in_millions ,
.33),quantile(budget_in_millions , .66)), degree =5)
+ poly(year_of_release, degree = 1)
+ poly(duration_in_hours, degree = 2)
+ poly(total_number_languages, degree = 1)
+bs(total_number_of_actors, knots = c(quantile(total_number_of_actors , .10),quantile(total_number_of_actors ,
.30), quantile(total_number_of_actors , .50)), degree =2)
+ poly(total_number_of_directors, degree = 3)
+ poly(total_number_of_producers, degree = 3)
+ bs(total_number_of_production_companies, knots =
c(quantile(total_number_of_production_companies , .33),quantile(total_number_of_production_companies ,
.66)), degree =1)
+ poly(qty_top_actors, degree = 2)
+ poly(qty_worst_actors, degree = 1) +
genre_action + genre_animation +
genre_biography + genre_comedy + genre_crime + genre_documentary +
genre_drama + genre_family + genre_fantasy + genre_filmnoir +
genre_history + genre_horror + genre_music + genre_romance +
genre_war + main_actor1_is_female + main_actor2_is_female +
main_actor3_is_female +
top_direct_flag + worst_direct_flag +
month_of_release_3 + month_of_release_6 + month_of_release_7 +
month_of_release_8 + month_of_release_9 + main_lang_deutsch +
main_lang_eesti + main_lang_english + main_lang_espanol +
main_lang_latin + main_lang_nederlands + main_lang_portugues +
main_lang_suomi + main_lang_tiang_viet +
main_lang_hangug_eo_joseonmal + main_lang_guang_zhou_hua_guang_zhou_hua +
main_production_country_argentina + main_production_country_belgium +
```

main\_production\_country\_bulgaria + main\_production\_country\_ecuador +  
main\_production\_country\_finland + main\_production\_country\_germany +  
main\_production\_country\_greece + main\_production\_country\_hungary +  
main\_production\_country\_iceland + main\_production\_country\_netherlands +  
main\_production\_country\_new\_zealand + main\_production\_country\_norway +  
main\_production\_country\_philippines + main\_production\_country\_singapore +  
main\_production\_country\_switzerland + main\_production\_country\_united\_kingdom)

**Appendix 6 - GENRE IMPLICATIONS:** Results of our model showing the estimates and significance levels of genres used as predictors.

genre_action	-0.232*** (0.034)	genre_fantasy	-0.124*** (0.044)
genre_animation	0.451*** (0.069)	genre_filmnoir	0.189 (0.176)
genre_biography	0.026 (0.056)	genre_history	-0.126* (0.067)
genre_comedy	-0.124*** (0.030)	genre_horror	-0.405*** (0.047)
genre_crime	0.007 (0.032)	genre_music	-0.027 (0.081)
genre_documentary	0.529*** (0.186)	genre_romance	-0.053 (0.036)
genre_drama	0.087*** (0.031)	genre_war	-0.094 (0.077)
genre_family	-0.201*** (0.054)		

**Appendix 7 - MAIN PRODUCTION COUNTRY IMPLICATIONS:** Results of our model showing the estimates and significance levels of main production countries used as predictors.

main_production_country_argentina	-1.075* (0.631)	main_production_country_iceland	1.132* (0.633)
main_production_country_belgium	0.722** (0.317)	main_production_country_netherlands	-0.443 (0.315)
main_production_country_bulgaria	-1.824*** (0.632)	main_production_country_new_zealand	0.651*** (0.180)
main_production_country_ecuador	-1.356** (0.631)	main_production_country_norway	-0.804* (0.446)
main_production_country_finland	-0.415 (0.446)	main_production_country_philippines	0.669 (0.631)
main_production_country_germany	0.120** (0.060)	main_production_country_singapore	-0.589 (0.447)
main_production_country_greece	0.573 (0.630)	main_production_country_switzerland	-1.186*** (0.366)
main_production_country_hungary	-0.534 (0.366)	main_production_country_united_kingdom	0.126*** (0.042)

**Appendix 8 - MAIN LANGUAGE IMPLICATIONS:** Results of our model showing the estimates and significance levels of the main language of a movie which were used as predictors in the final model.

main_lang_deutsch	0.023 (0.110)	main_lang_portugues	0.605* (0.319)
main_lang_eesti	0.724 (0.631)	main_lang_suomi	1.319* (0.777)
main_lang_english	-0.083* (0.046)	main_lang_tiang_viet	-0.893 (0.632)
main_lang_espanol	-0.175 (0.110)	main_lang_hangug_eo_joseonmal	0.793 (0.631)
main_lang_latin	0.232 (0.243)	main_lang_guang_zhou_hua_guang_zhou_hua	0.301 (0.204)
main_lang_nederlands	1.253**		

**Appendix 9 - MAIN ACTOR IS FEMALE IMPLICATIONS:** Results of our model showing the estimates and significance levels of the main actors are female predictors.

main_actor1_is_female	-0.180*** (0.030)
main_actor2_is_female	-0.063** (0.025)

#### **Works Cited:**

Statista. (2021, August 12). *Distribution of lead actors in movies in the United States from 2011 to 2020, by gender.*

Retrieved from Statista:

<https://www.statista.com/statistics/692465/distribution-lead-actors-gender/>