MGSC 661 Midterm Project:

THE 2021 IMDb PREDICTION CHALLENGE

R-ockets, R-ock it!

I. Introduction

Scroll… "overrated" …scroll… "maybe" …scroll… "already seen it" … we've all been there, spending half an hour scrolling aimlessly through pages of movies on the Netflix account that belongs to your old roommate while your group of friends or family squabble over which movie to watch. But what if it didn't have to be this way? What if there was a way to predict the rating of a movie almost instantly without having to listen the opinions of your loved ones? Sound too good to be true? It's not. This is exactly what our model is designed to do. Just name a movie, and our model will predict its IMDB rating. No more family feuds or petty arguments, just the art of predictive analytics!

Our team analysed a dataset of 2957 movies to learn how to design the most effective movie rating predicative model we could. For each of the observations (movies), the dataset contains one dependent variable (the variable we want to predict) and 47 independent variables (the characteristics of movies). The dependent variable is IMDB movie rating, and the independent variables are characteristics like the budget of the movie, the year it was released, or the name of the lead actor.

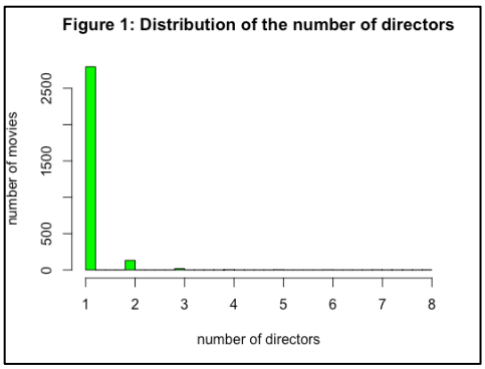
We began our analysis by visualizing the predictor variables in the dataset to investigate how characteristics varied across our dataset. Next, we tested the individual variables in our dataset for issues like correlation, outliers, and collinearity. Finally, we investigated the relationships between our dependent variable (Y) and each of the predators (x_i). We ran simple linear regressions between Y and each x_i to check for heteroskedasticity and examined the p-values and R-squared scores to assess which predictors had linear predictive power. The goal was to know our dataset inside and out so that we could take an evidence-based approach to designing our model. This report provides an overview of our data analysis process, model design, results, and a summary of our learnings about what makes a good movie. It was challenging work, but worth it. After all, who doesn't love statistics and movies?

II. Data Description

Before designing a model, we conducted data exploration to understand the distribution of the independent and dependent variables and analyse the relationship that each of the independent variables has with the dependent variable. The process and findings of our data exploration are below.

1. Independent variables

We assessed the distribution of the independent variables by visualizing them and measuring their skewness. We generated histograms and boxplots for the numerical independent variables (see

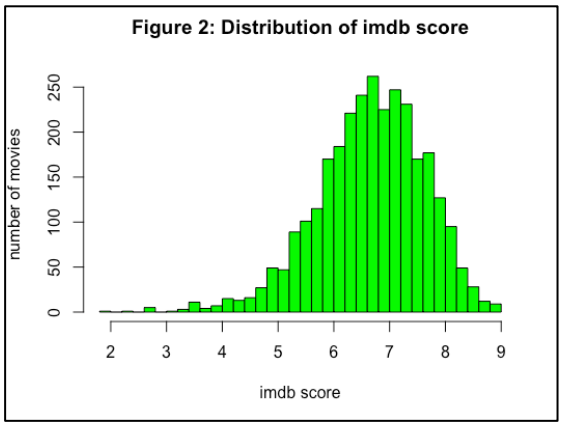


Appendix 3-1), and a bar chart for the binary independent variables (i.e., the genres of the movies) (see Appendix 3-2). These visualizations provided preliminary insights about the data, like that very few leading actors are female. They also demonstrated that some of the independent variables were at least 90% or more unary. This means that at least 90% of the observations for these variables have the same value. The number of directors variable is unary because there is one director 95% of the time (Figure 1). The language of films is nearly unary because it is English 89% of the time (see Appendix 1). The production country of the movies looks close to unary too, but in fact, the movies are only produced in the U.S.A. 71% of the time (see Appendix 2). Finally, we made pie charts of the binary independent variables. These indicated that there are no movies in the data set from the reality tv or short film genre. Based on this analysis, we removed the reality tv and short films variables and the number of directors variable. Then, we converted the main language and production country variables into binary variables because even though they were not unary, most observations had the same value, and we needed some way to transform them into numeric or dummy variables so they could be included in our regressions. Main language became 1 if the language was English (0 otherwise), and production company became 1 if the country was the U.S.A. (0 otherwise).

After assessing the visualizations of the independent variables, we tested the skewness of each independent variable (results in Appendix 4). These levels of skewness are valuable because they efficiently indicate how the observations for each independent variable are distributed. For instance, the documentary genre binary variable is very right skewed (skewness = 14.5), which indicates the for most of the observations, the value for this variable is 0. Thus, there are very few documentary films in our dataset.

Next, we tested for correlation amongst our numeric independent variables. We excluded the binary and character independent variables and ran three correlation tests (Pearson, Kendall, and Spearman) on the remaining numeric independent variables, then we created correlation heat maps for each test. All the correlation matrices and heat maps indicated that there are only weak (under 0.5) levels of correlation between any given pair of independent variables (see Appendix 5). Therefore, there is no evidence for removing any of the independent variables because of their correlation with another independent variable.

2. Dependent variable



To investigate our dependent variable (IMDB score) we first created a histogram to visualize its distribution (Figure 2). The distribution is slightly left skewed (skewness = -0.59), which indicates that more of the movies have a score that is lower than the mean score.

To conclude our exploration of the variables independently, we tested the dataset for outliers and collinearity. To detect outliers, we ran a multiple linear regression of our dependent variable on our numeric independent variables. The Bonferroni test of this regression identified 6 observations as outliers (observations 633, 895, 2045, 2310, 2718, and 526). We removed these observations, ran the regression again, and conducted another Bonferroni test. This second test identified another outlier, but when we removed this outlier, it did not significantly affect the results of our regressions, so we decided to only remove the 6 observations identified by the first test. Finally, we ran a variance inflation factor (VIF) test on the multiple linear regression (using the dataset without outliers) to test for collinearity. All the independent variables had a VIF score under 4 (see Appendix 6); thus, we concluded that there was no evidence of collinearity amongst the variables.

3. Relationship between independent variables and the dependent variable

We split the numeric independent variables (predictors) into two major parts based on their data type, namely numeric and character variables. We examined the relationship between the dependent variable and each numeric independent variable by conducting a series of visual and numerical tests. The character independent variables were analyzed separately (next section). First, we obtained the correlation coefficients between the dependent variable and each independent variable to get the sense and strength of each of relationship. We also visualized each relationship with scatter plots. Then, we ran simple linear regressions between the dependent variable and each independent variable to examine their relationship in detail. We recorded the coefficient, the corresponding p-value, and the R-squared value of each simple linear regression. Then, we conducted non-constant variance (NCV) tests to check if heteroskedasticity is presents in any of the independent variables because heteroskedasticity can cause biased standard errors of the linear regression coefficient and the p-value might not tell the true significance of the relationship. The independent variables with heteroskedasticity (NCV test p-value < 0.05) are marked red in Appendix 7 in the

appendices. We corrected the heteroskedasticity and obtained the unbiased p-value for linear regression of those variables. Then, we tested the linearity of each predictor by running residual plots on each of the simple linear regressions. The non-linear predictors (with linearity test p-value < 0.05) are marked green in Appendix 7. There are 37 numeric predictors, 5 of which are identified as non-linear and 32 of which are identified as linear. The statistics of all the tests we conducted are recorded in Appendix 7.

III. Model Selection

Before building models, we analyzed the information we obtained in the previous steps. In previous sections, we have split the predictors into numeric and character predictors. For numeric predictors, we further split them into linear (linearity test p-value > 0.05) and non-linear (linearity test p-value < 0.05) and then analyzed them separately because some of the statistics we obtained in the previous step, like the p-value of simple linear regression coefficient and the R-squared value, are not good performance indicators of the non-linear predictors. The binary dummy variables are identified as linear numeric predictors as well.

1. Analyzing non-linear numeric-type variables

We tried to fit the relationship between the dependent variable (imdb_score) and each non-linear predictor with polynomial regressions of different degrees (2 to 5), and splines regressions with different numbers of knots (1 to 5) and different degrees (1 to 5). We ran 10-fold cross validation tests of each regression and recorded the hyperparameters that gave the lowest MSE. Our goal was to get a general idea of what regression we should try for these non-linear predictors when building the model, but the results did not necessarily show us the final hyperparameters to use in our model because these regressions are only between the dependent variable and each predictor.

2. Analyzing linear numeric-type variables

For all the linear predictors, we decided to fit them with linear regressions to avoid overfitting, even though fitting polynomial regressions have higher R-squared values. There are 32 linear predictors, but this is too many to include in our model. To further narrow down the set of predictors to be selected and the priority of adding them to our model, we ranked the linear predictors based on their unbiased p-value (after correcting heteroskedasticity) and R-squared value of the simple linear regression between each of them and the dependent variable as the p-value indicates the significance of the relationship between them and R-squared value reflects the linear predictive power of the

predictor. The lower the simple linear regression p-value and the higher the R-squared, the higher the priority to be considered for our model. The predictors in Appendix 7 are sorted by their unbiased p-value from smallest to largest. Those linear predictors with unbiased p-value > 0.05 were marked yellow in Appendix 7 and will not be added to the model since their relationship with the dependent variable is insignificant.

3. Analyzing and processing character-type variables

We decided to include character-type variables such as directors' names and actors' names in our model (logically, these people have a huge impact on ratings!), but to do so we had to resolve the issue of having more than 1100 levels in each of these predictors. We wanted to find actors/directors who have played in/directed movies more than a certain number and who have a significant impact on the ratings of those movies. Through a feature engineering process, we identified those who played in/directed more than 6 movies (chosen based on the mean and median of the frequency) and among those, we again chose the ones who significantly impacted the ratings (having a lower p-value than 0.05 when run in a solo regression against ratings). We assigned the "Others" label to all other actors/directors, thus resolving the issue of 1100 levels, and we included the significant actors/directors in our model. We didn't include second and third actors and producers since they didn't significantly contribute to the ratings.

4. Building and testing the model

After analyzing and processing different types of predictors and determining their relationship between the dependent variable, we started to build our predictive model. We decided whether to include certain predictors or not based on the adjusted R-squared value of the model since it indicates whether the predictor improves the model or not. We included a predictor if the adjusted R-squared value of the model increased.

We started with adding the non-binary numeric predictors. There are 9 non-binary predictors (both linear and non-linear, marked blue in Appendix 7) out of 37 numeric predictors. For linear predictors, we just added them to the model one by one while checking the R-squared each time. Then, we tried the polynomial regressions on the non-linear predictors. We ran polynomial ANOVA test on each of them to determine the optimal degree. We then added them one by one, checking the adjusted R-squared of the model each time. We tried spline regressions on them based on the prior information we obtained in section 3, but it turns out that spline regressions generally do not perform well on the

dataset since adjusted R-squared of model was lower and the MSE was higher than when we applied polynomial regressions. We verified that by looking at the distribution of the each of these predictors on scatter plots and we concluded spline regressions are not suitable for fitting our data.

For those binary numeric predictors (dummy variables), we decided not to fit them with non-linear regressions to avoid overfitting since their values are either 0 or 1. Instead, we just applied simple linear regressions on each of them and added them one by one to the model while observing the adjusted R-squared of each model. We did the same for character-type predictors like main_actor1_name and main_director_name. We also tried to add interaction terms by multiplying two predictors in our model using our intuition. The logic of each interaction term (A*B) is that the effect of one predictor (A) on the ratings is different for different values of the other predictor (B). For example, we reasoned that the effect that budget had on the rating would be different for different years of release. It turns out that adding certain interaction terms increased the adjusted R-squared, and reduced the MSE, which was calculated using both the validation set test method and the LOOCV method.

After getting our base model, we tweaked it by varying the polynomial degrees by +/- 1, trying different reasonable interaction terms, and adding/removing certain predictors. We tested each model primarily with LOOCV tests because there is no randomness in the result so that we could compare MSEs across different models. But we also tried 10-fold cross-validation and validation set test (with 80% training) on the models that we were particularly interested in. The objective of our model selection process was to get the lowest MSE possible using the LOOCV test. Here is the final model we selected:

```
mreg <- lm(imdb_score~
        poly(duration_in_hours,4)+
        poly(year_of_release,3)+
        poly(total_number_of_actors,2)+
        month_of_release+
        total_number_languages+
        poly(budget_in_millions,3)+
        poly(total_number_of_producers,4)+
        main_production_country_US+
        genre_action*genre_comedy+
        genre_drama*genre_horror+
        main_actor1_is_female*main_production_country_US+
        budget_in_millions*year_of_release+
        duration_in_hours*genre_comedy+
        main_actor1_name + main_director_name)
```

IV. Managerial Implications

Making a movie that will get a high rating is a careful balancing act. Exploring the data, designing a model, and analysing our model's results taught us that one must consider both the characteristics of a movie (e.g., genre, duration, month of release) and the people who will work on the movie (i.e., director and actors). There is no perfect recipe, but we can make the following recommendations about what makes a good movie:

1. Recommendations: movie characteristics

When it comes to movie characteristics, a few elements tend to significantly affect a movie's ratings. First, as one would expect, the main actor of a movie is a significant factor contributing to the rating of a movie. How many times have you thought about something like "Oh Leonardo Di Caprio plays in this movie; it must be a good one!". Similarly, the main director is also a critical factor to take into consideration. The director is responsible for the filmmaking, orchestrating the screenplay and shooting of the movie, it is therefore logical that the director is a strong contributor to a movie's success.

Moreover, the budget allocated to a movie generally correlates with higher scores. Although some very low-budget movies have been absolute blockbusters, a higher budget tends to correlate with a better movie rating: a high budget allows to have high-profile actors who can be paid accordingly, more resources for better screenplay, better CGI, better camera plays, etc.. As such, since the budget determines the extent and the quality of the available resources, it generally tends to have a noticeable impact on a movie's success.

In addition, we can also consider the duration of the movie as a predictor of its success. Some of the most successful movies of all times share the common characteristic of having a longer than average runtime. Among the most notables, Titanic has a runtime of 3.5 hours, The Irishman has a runtime close to 4 hours, The Godfather movies are close to 3 hours in duration each, for a trilogy totaling close to 8 hours, and Avengers: Endgame has a runtime slightly above 3 hours. Generally, longer movies tend to obtain higher ratings. A longer movie can be an indicator of a well-written, compelling storyline, as well as potentially indicating elaborate and complex character development. However, it is worth noting that the impact of a movie's duration on the score is weaker than the impact that the budget and main actor & director have.

Finally, the genre of the movie is another important element to take into consideration when listing the factors that affect a movie's quality. Oftentimes, genres have a target audience, and some movie genres are inherently easier to produce than others.

To sum up, according to the results of our model, the main actor, the main director, the allocated budget, the duration as well as the genre of the movie are among the most important characteristics for determining a movie's success.

2. Recommendations: people to include (or not include) in a movie

Movie stars matter! The results of our model indicated that the lead actor of a film has a significant impact, good or bad, on the rating of a movie. Out of the lead actors in our dataset, Jim Carey, Matt Damon, Will Farrell, Johnny Depp, and Tom Hanks have the highest positive impact on movie rating. We recommend casting these actors in the lead role. On the other hand, there are some actors that should not be cast in the lead role. Out of the actors in our dataset, Chris Lambert, Jean-Claude Van Damme, Steven Seagal, and Brendan Fraser have the highest negative impact on a movie's rating. We do not recommend casting these actors in leading roles. Figure 3 below shows how much these lead actors are expected to affect movie rating. For example, our model predicts that if everything else is held constant, casting Matt Damon as the lead actor will increase movie rating by 0.437.

Figure 3 - Impact of lead actors on movie rating

	Dependent variable:
	IMDB Score
Brendan Fraser	-0.511*
	(0.262)
Christopher Lambert	-1.142***
	(0.281)
Jean-Claude Van Damme	-0.717***
	(0.276)
Jim Carrey	0.469**
	(0.200)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	IMDB Score
Johnny Depp	0.341**
	(0.158)
Matt Damon	0.437**
	(0.187)
Steven Seagal	-0.529**
	(0.263)
Tom Hanks	0.325**
	(0.160)
Will Ferrell	0.432**
	(0.207)

Choosing the lead actor is important but selecting the right director can is more consequential (the average impact on rating, good or bad, of directors is higher than the impact of actors). Our model indicated that directors have a significant impact, good or bad, on a movie's rating. Out of the directors in our dataset, Christopher Nolan, Danny Boyle, David Fincher, James Cameron, Joel Coen, Peter Jackson, Robert Zemeckis, Steven Spielberg, Wes Anderson, and Woody Allen have the highest

positive impact on movie rating. We recommend selecting one of these directors. Figure 4 below shows how much these directors are expected to affect movie rating. For example, our model predicts that if everything else is held constant, selecting Christopher Nolan to direct a movie will increase its rating by 1.169. Our model did not identify any directors who have a significant negative impact on movie rating. Therefore, there are no specific directors that we can recommend avoiding.

Figure 4 - Impact of directors on movie rating

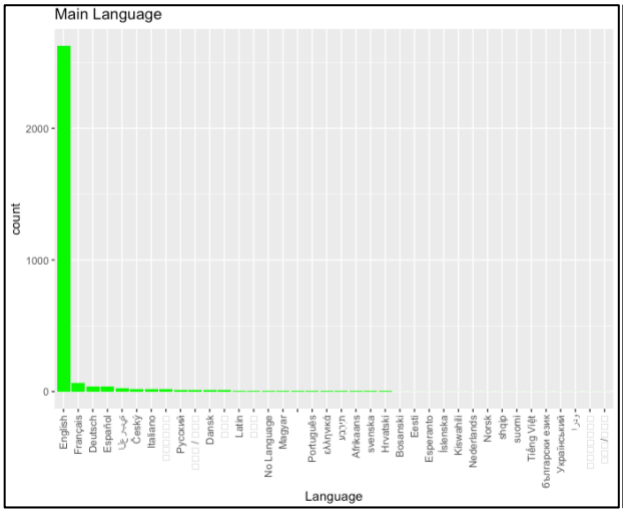
	Dependent variable:
	IMDB Score
Christopher Nolan	1.169***
	(0.300)
Danny Boyle	0.632**
	(0.271)
David Fincher	0.618**
	(0.253)
James Cameron	1.091***
	(0.297)
Joel Coen	0.845***
	(0.249)
Note:	*p<0.1; **p<0.05; ***p<0.01

	Dependent variable:
	IMDB Score
Peter Jackson	1.019***
	(0.273)
Robert Zemeckis	0.562**
	(0.221)
Steven Spielberg	0.459***
	(0.158)
Wes Anderson	0.964***
	(0.269)
Woody Allen	0.584***
	(0.220)

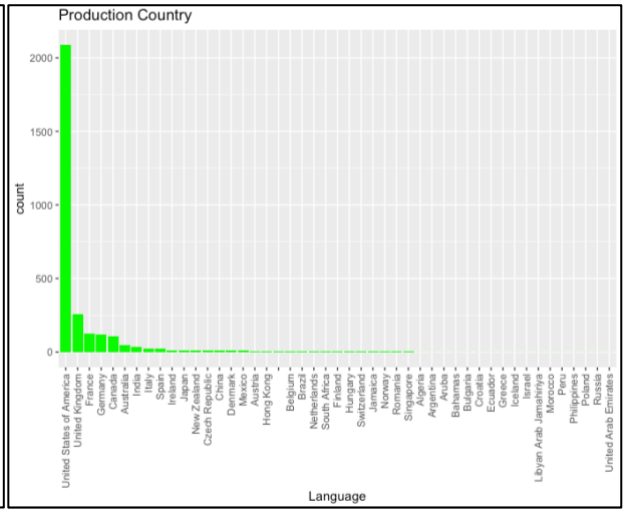
It needs to be acknowledged that all these lead actors and directors are men. This is indicative of the fact that most of the lead actors and directors in our dataset were male (78% of lead actors and 96% of directors were male), and therefore our model was more likely to recommend male leads and directors because it does not have a larger enough sample of female leads and directors. If we lower our standard for what constitutes a significant relationship (decided that a p-value < 0.05, instead of 0.01, constitutes a statistically significant relationship), we can say that Meryl Streep and Cate Blanchett have a significant positive impact on movie rating. Nevertheless, in the future, as more women get lead roles and direct movies, we strongly recommend conducting further research to determine their impact on ratings.

V. Appendices

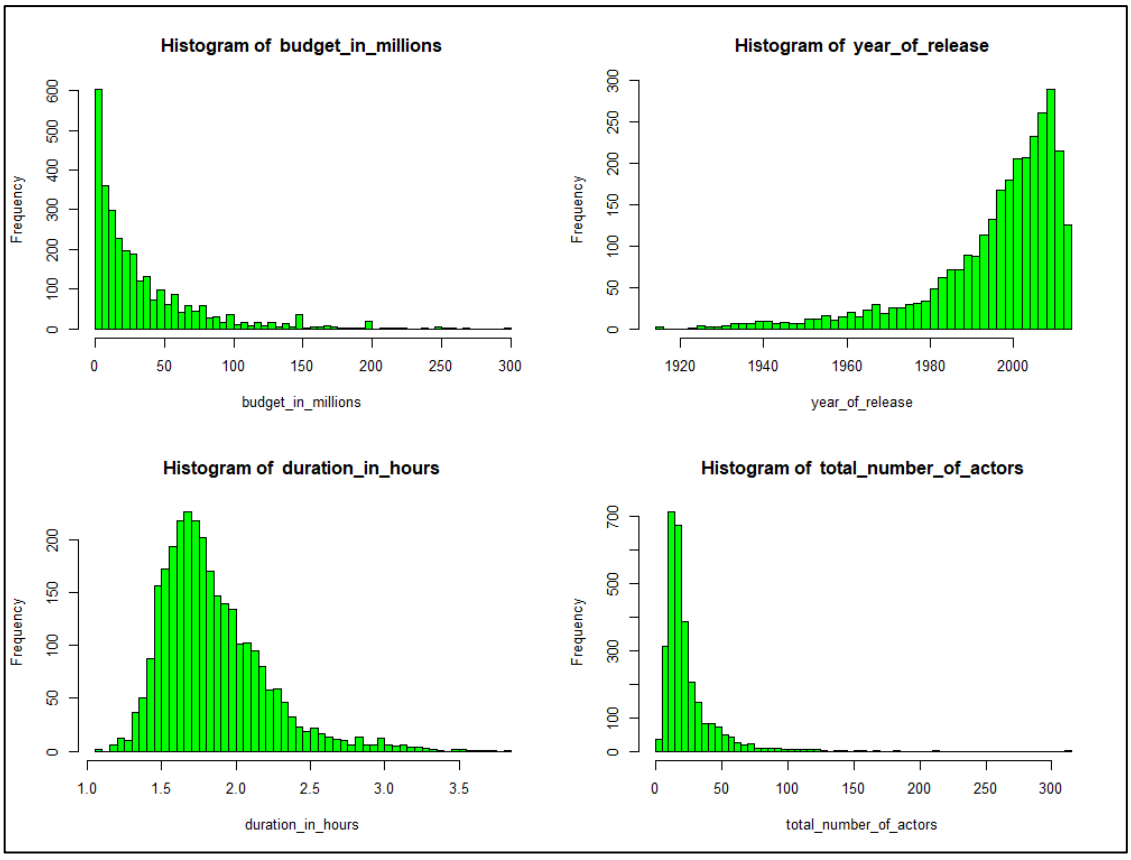
Appendix 1

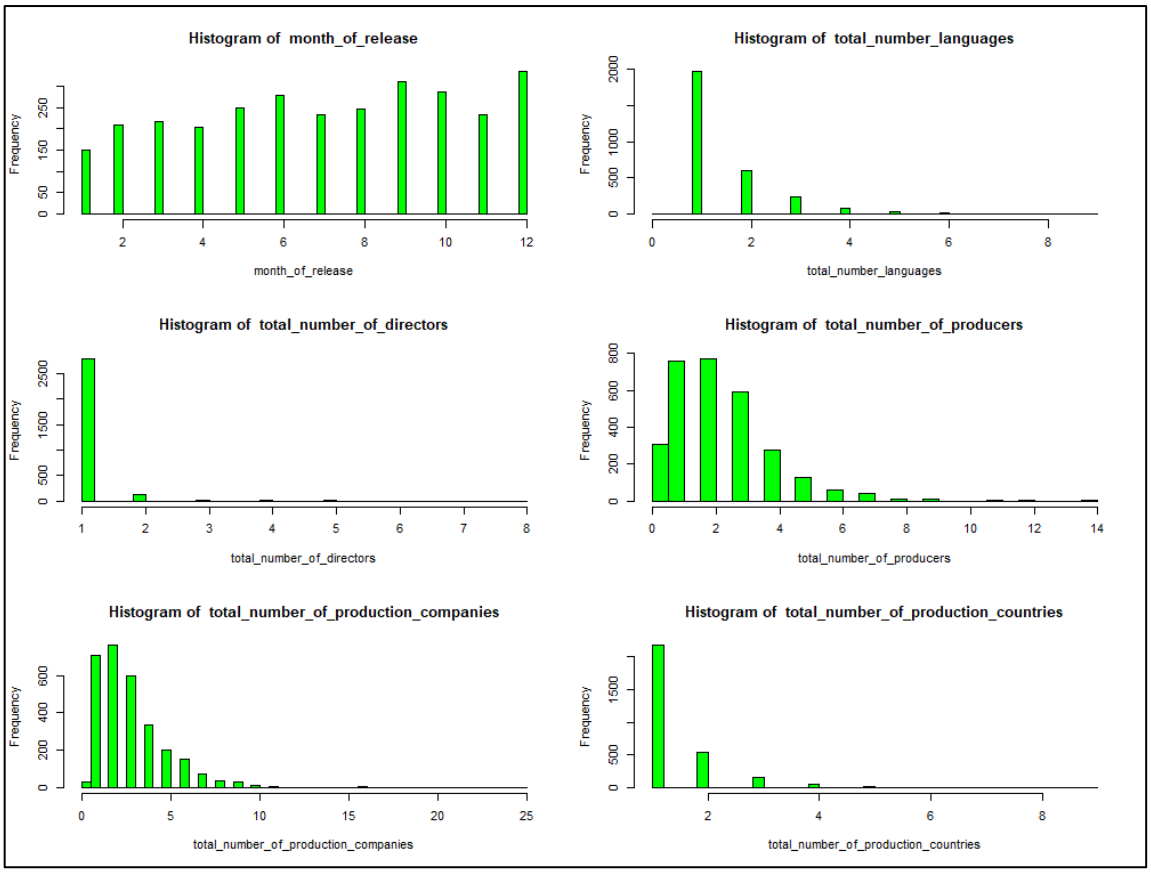


Appendix 2

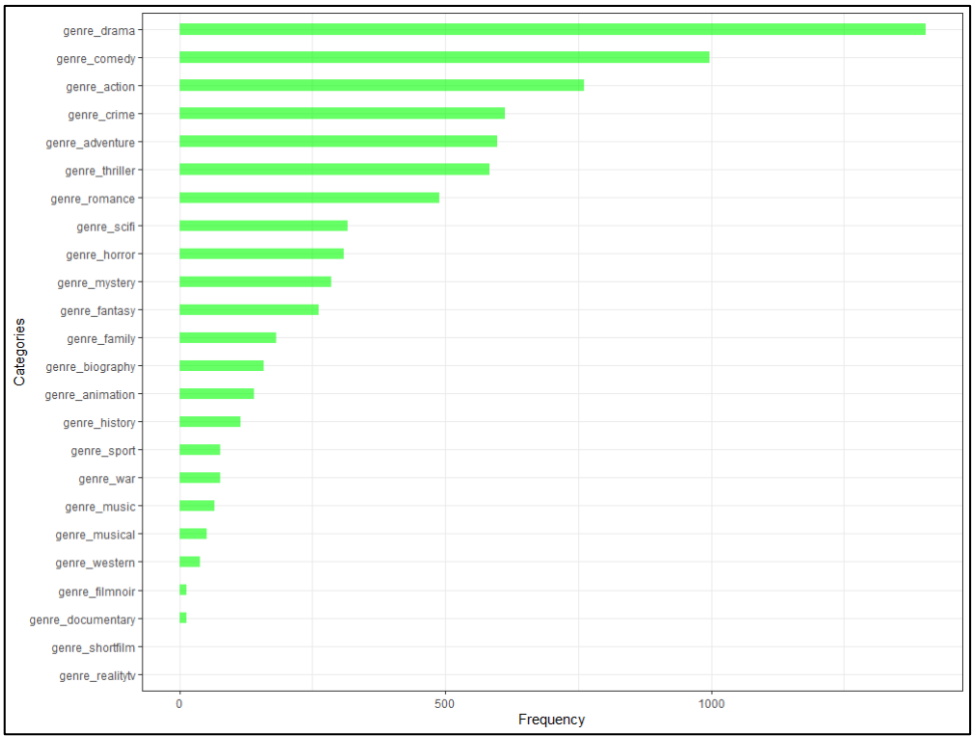


Appendix 3-1 – Histograms of numerical independent variables

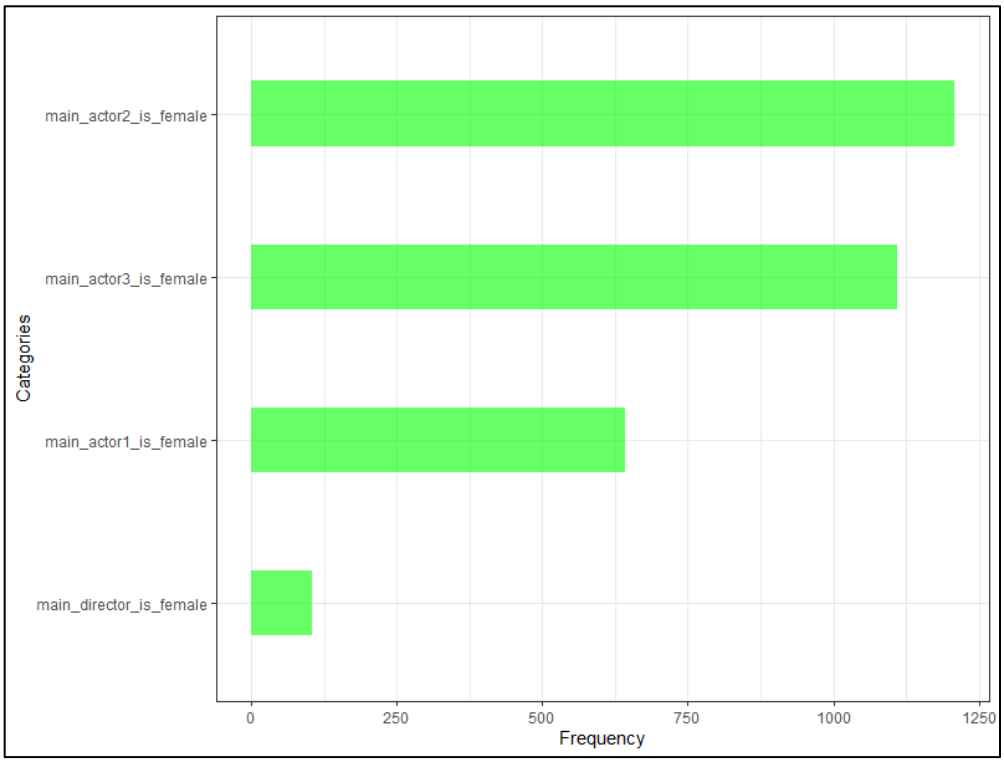




Appendix 3-2 – Bar chart for Binary variables (genres)



Appendix 3-2 Continued – Bar chart for Binary variables (non-genre binary variables)

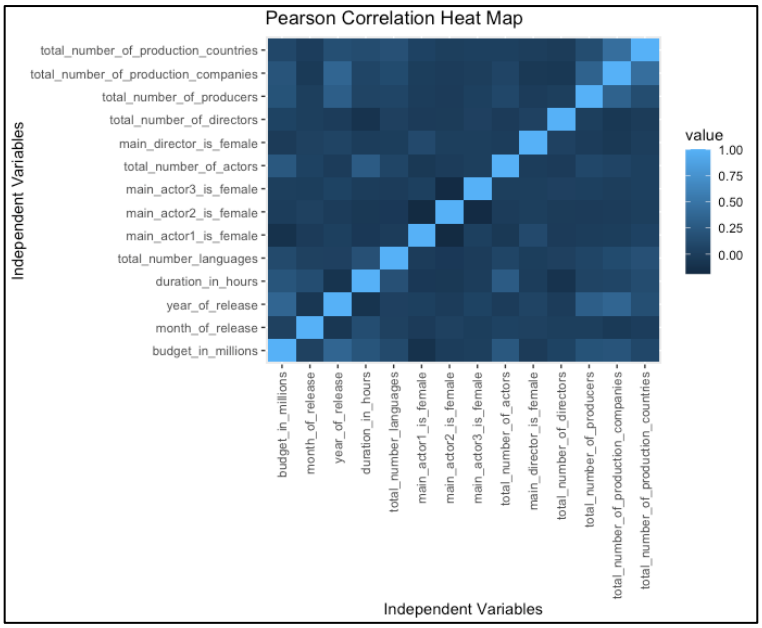


Appendix 4 – Skewness of independent variables (predictors)

Independent variables	Skewness	Independent variables	Skewness
budget_in_millions	2.225873851	genre_history	4.766422386
month_of_release	-0.153748199	genre_horror	2.583312348
year_of_release	-1.748849914	genre_music	4.766422386
duration_in_hours	1.421702863	genre_musical	6.462606699
total_number_languages	2.374079455	genre_mystery	7.410771544
genre_action	1.105945063	genre_romance	2.732805353
genre_adventure	1.483170336	genre_scifi	1.802554052
genre_animation	4.241863347	genre_sport	2.536871179
genre_biography	3.953382691	genre_thriller	5.947892286
genre_comedy	0.685130476	genre_war	1.517571866
genre_crime	1.441963182	genre_western	5.990131849
genre_documentary	14.97191597	total_number_of_actors	8.644279168
genre_drama	0.099683356	total_number_of_directors	3.71589129
genre_family	3.645676204	total_number_of_producers	1.191503076
genre_fantasy	2.892810734	total_number_of_production_companies	2.116388226
genre_filmnoir	14.41989372	total_number_of_production_countries	2.674525591

Note: predictors genre_shortfilm and genre_realitytv are excluded from Appendix 4 because every observation is 0 for these variables.

Appendix 5 – Pearson Correlation Heat Map



Appendix 6 – Results of VIF test for collinearity

independent variable	VIF Score	independent variable	VIF Score
budget_in_millions	1.84	genre_history	1.23
month_of_release	1.05	genre_horror	1.56
year_of_release	1.61	genre_music	1.08
duration_in_hours	1.70	genre_musical	1.07
total_number_languages	1.10	genre_mystery	1.22
genre_action	1.70	genre_romance	1.37
genre_adventure	1.70	genre_scifi	1.32
genre_animation	1.57	genre_sport	1.10
genre_biography	1.26	genre_thriller	1.57
genre_comedy	2.10	genre_war	1.13
genre_crime	1.42	genre_western	1.09
genre_documentary	1.06	total_number_of_actors	1.17
genre_drama	2.09	total_number_of_directors	1.20
genre_family	1.25	total_number_of_producers	1.22
genre_fantasy	1.28	total_number_of_production_companies	1.52
genre_filmnoir	1.08	total_number_of_production_countries	1.30

Appendix 7 – Test statistics of relationship between dependent variable and each predictor (numeric predictors only)

Predictors	corr_coef	linear_reg_coef	p-value of coef	p-value of coef no hetero	r-squared	residual Pr-value	NCV p-value
duration_in_hours	0.361797597	0.964557786	7.76E-92	1.50E-86	0.130897501	0.000417141	3.79E-08
genre_drama	0.295821689	0.561598435	1.35E-60	1.18E-61	0.087510472	0.47388097	1.21E-13
year_of_release	-0.283478314	-0.016486351	1.38E-55	1.37E-81	0.080359954	6.32E-05	1.31E-07
genre_horror	-0.217405827	-0.674758538	7.30E-33	8.31E-28	0.047265294	0.741459074	0.003328788
total_number_of_actors	0.215274051	0.010153961	3.07E-32	1.47E-18	0.046342917	8.89E-08	1.07E-05
genre_comedy	-0.192895975	-0.386740877	4.26E-26	4.26E-26	0.037208857	0.795182899	0.898917647
genre_action	-0.156160528	-0.338300173	1.51E-17	1.51E-17	0.02438611	0.502941239	0.502557731
genre_biography	0.148067223	0.621387257	6.52E-16	4.33E-29	0.021923902	0.664450744	2.48E-06
month_of_release	0.131321531	0.037028201	8.25E-13	8.25E-13	0.017245345	0.05659867	0.268759738
genre_history	0.123133546	0.602877671	1.98E-11	1.09E-20	0.01516187	0.673128731	8.75E-05
genre_filmnoir	0.097121308	1.339191954	1.27E-07	1.22E-85	0.009432548	0.687128984	0.013030463
main_actor1_is_female	-0.09608581	-0.220701171	1.73E-07	1.73E-07	0.009232483	0.729448902	0.715357455
genre_war	0.087596153	0.523974317	1.91E-06	9.58E-10	0.007673086	0.681863596	0.011226704
genre_fantasy	-0.085005359	-0.283186063	3.82E-06	4.26E-05	0.007225911	0.711644988	0.000247239
genre_family	-0.084932572	-0.334533315	3.89E-06	3.89E-06	0.007213542	0.708438987	0.961747505
total_number_languages	0.078185284	0.076475811	2.15E-05	2.15E-05	0.006112939	0.416267204	0.267577147
genre_scifi	-0.077707056	-0.238791333	2.41E-05	2.41E-05	0.006038387	0.712981941	0.042530985
main_lang_English	-0.071426879	-0.215043967	0.000104108	0.000104108	0.005101799	0.674610596	0.147515592
budget_in_millions	-0.071390414	-0.001632563	0.000104962	0.000104962	0.005096591	2.70E-25	0.914234343
main_production_country_US	-0.065469982	-0.136268538	0.000375929	0.0002651	0.004286319	0.659702079	0.036598174
genre_documentary	0.058739573	0.840383304	0.001421967	0.001421967	0.003450337	0.689977343	0.096576014
genre_musical	0.057699171	0.419501544	0.001727025	1.46E-05	0.003329194	0.685961969	0.013129312
genre_western	0.055030824	0.46247761	0.002804278	0.002804278	0.003028392	0.688358061	0.366025276
genre_crime	0.04026489	0.094171146	0.028830966	0.02184438	0.001621261	0.676576903	0.015049029
genre_thriller	-0.038631129	-0.091885605	0.035990629	0.035990629	0.001492364	0.707182328	0.300359592
main_actor3_is_female	-0.03814981	-0.074662446	0.03836941	0.03836941	0.001455408	0.716568869	0.627194387
main_actor2_is_female	-0.033017751	-0.063700947	0.073110707	0.073110707	0.001090172	0.71300543	0.342192071
genre_adventure	-0.031566974	-0.074466374	0.086647531	0.086647531	0.000996474	0.64645061	0.637512155
total_number_of_production_companies	-0.02778048	-0.012828419	0.13161851	0.13161851	0.000771755	0.088529957	0.868845882
genre_sport	0.02732236	0.162397846	0.13810714	0.13810714	0.000746511	0.689948678	0.373767925
genre_music	0.019943026	0.128746597	0.27912768	0.27912768	0.000397724	0.691222754	0.031982426
genre_animation	0.019583267	0.08699216	0.287893721	0.1905763	0.000383504	0.757794928	0.002018056
main_director_is_female	-0.015683509	-0.080590993	0.394719087	0.394719087	0.000245972	0.695335464	0.231284259
genre_romance	0.014712308	0.037588521	0.42464952	0.3788324	0.000216452	0.687330282	0.000168234
genre_mystery	0.010421599	0.033431531	0.571716605	0.571716605	0.00010861	0.690622597	0.612765047
total_number_of_producers	0.003598814	0.002102467	0.845170644	0.8437032	1.30E-05	0.003083463	0.00563803
total_number_of_production_countries	0.000389488	0.000489446	0.983138071	0.983138071	1.52E-07	0.348017795	0.459531916

VI. Code

All the code used to visualize and process the dataset and create and test our model is in the file: "Final_Group1_MGSC661_Midterm", which was included with our submission.