BROUGHT TO YOU BY:

MGSC661 MULTIVARIATE STATISTICAL ANALYSIS

JUAN CAMILO SERPA

THE 2023 IMBD MOVIE PREDICTION CHALLENGE

1. INTRODUCTION

In 2023, the global movies industry market size was estimated to be around 90.92 billion dollars. While also being a financially important industry it also plays a huge role in popular culture. Within the global movie industry is a sector that focuses on the reviewing of movies. One of the leaders within this industry is a website known as IMDB. IMDB was initially founded in 1990 with its function focusing on the review and critique of movies. Over the years it has grown exponentially and has expanded to being one of the largest movie critics in the world. Although it may seem that movie ratings may not have a sway or an effect on the performance of movies. The scale and reach of IMDB has changed this narrative. Individuals may or may not decide to go to a movie based upon the IMDB score that it receives. Currently IMDB reviews are conducted by IMDB users who all have their own biases and prejudice when reviewing a movie. For example, say an individual is not a fan of Christopher Nolan as a person and decides to provide a low review for his movie not based on the actual movie but rather because of their opinion of him as a person. This would be an unjust rating that could potentially have an effect on the amount of viewership that is given to the movie. As a result of this it is imperative that a model be built that helps determine the ratings for movies based on unbiased and accurate measures.

2. DATA DESCRIPTION

EXPLORATORY ANALYSIS OF THE DATA

Columns within the dataset can be split up within 5 categories; Identifiers, Dependent Variable, Film Characteristics, Cast Characteristics and Production Characteristics.

EXAMINING THE 5 CATEGORIES OF VARIABLES

1. IDENTIFIERS

The variables of movie_title, movie_id and imdb_link are used to identify specific movies within the dataset. These identifiers were deemed as unnecessary and were removed from the dataset for our analysis as they do not provide any predictive power and in inadvertent cases may result in overfitting or inaccuracy of the model.

2. DEPENDENT VARIABLE

2. DATA DESCRIPTION

EXAMINING THE 5 CATEGORIES OF VARIABLES (CONTINUED)

The dependent variable within the dataset that will be used to predict within our model is imdb_score. All other variables will be assessed upon there relation to this variable.

3. FILM CHARACTERISTICS

The variables that fall under the category of film characteristics have both numerical and categorical information including but not limited to variables such as movie budget, release day, release month duration and language. These variables are in relation to the factors outside of the direct production of the film. While many of these variables appear to have significance in terms of their relationship to the quality of a movie, further data pre-processing has to be done so that they can effectively be implemented in our model.

4. CAST CHARACTERISTICS

The cast characteristics variables include information in regards to the actors associated with the movie. Not only are the actors names provided but also the IMDb star meter associated with them is also provided. The IMDb star meter is a metric used by IMDb to determine an actor's popularity. It is determined based on the amount of searches and page views that an actor or actress receives on the IMDb web page. In order to use the actor names within our analysis further data pre-processing will be required.

5. PRODUCTION CHARACTERISTICS

The final category of variables is within the domain of production. This category includes variables that highlight the areas of production associated with a movie such as the director, cinematographer and production company that worked on a movie. While these variables certainly play a role in the quality of a movie they are all stored in free text form meaning that further processing of the variables will be required in order to include this information in our model.

2. DATA DESCRIPTION

PRE-PROCESSING THE DATA

1. DUMMIFICATION OF CATEGORICAL COLUMNS

Categorical columns such as release month, language, country, maturity rating, distributor, director, actor 1, actor 2, actor 3, color of the film, ,cinematographer and production company were dummified. This was to ensure that there could be a possibility that these variables would be included within our model. Dummification is a technique that is used to apply binary values v(0 or 1) to a categorical variable so that it can be used for analysis and can be applied to predictive models. Although dummyfying these variables increased the dimensionality of our dataset, we deemed it necessary to ensure that key variables, potentially possessing high predictive power, were not excluded from our analysis. Additionally, the initial dummified genre categories were removed and replaced with new dummified versions of the genres listed in the genres columns. This was done because it was found within the exploratory phase that many types of genres such as family movies had been missing within the initial dummified versions.

2. ANALYSIS OF DATA DISTRIBUTION WITHIN THE DATASET

We conducted an analysis of the dataset variables using box and whisker plots and histograms. Box and whisker plots helped identify outliers, while histograms aided in understanding data distribution and potential skewness.

Numerical variables such as IMDb score, movie budget, duration, aspect ratio, number of news articles, and number of faces were analyzed using box and whisker plots. These variables were chosen due to their data spread, unlike categorical variables which have binary outputs. The same numerical variables were also plotted in histograms. (See Appendix 1 and Appendix 2)

3. REMOVAL OF OUTLIERS FOR EACH COLUMN

To remove outliers, we employed the interquartile range (IQR) method. This method identifies outliers as values outside the range of Q1 - 1.5 * IQR to Q3 + 1.5 * IQR. In a dataset, Q1 represents the median of the lower half, while Q3 denotes the median of the upper half of the data values. The IQR is a measure of which the middle 50% of the data lies. Using this formula to analyze each variable helps determine the normal range of data. Data points that fall outside this range are considered outliers and are filtered out accordingly.The IQR method was applied to the

2. DATA DESCRIPTION

PRE-PROCESSING THE DATA (CONTINUED)

categorical columns .The Removal of outliers is important as they have a low probability of occurring and are not representative of the population. This means it will only negatively affect the predictive power of the model that we plan to generate.

4. IDENTIFYING CORRELATIONS WITH IMDB SCORE

Another step of our exploration of the data included determining the top 10 positive and negative correlators within the dataset. This was done in order to get a deeper understanding into the data while identifying relationships that are occurring. It was determined that the strongest positive correlations occurred between IMDb Score and the movie's duration(.3964) , along with if the movie was a drama(.3455) and with the higher amount of news articles that it had (.245). The strongest negative correlations occurred when the movie was a horror (-.192), if the movie was released in a more recent year (-.1639) and if the movie was an action movie (-.1542). It should be noted that none of these correlations are particularly strong. (See Appendix 3 and 4) Correlations between numerical variables and IMDB score were also explored by plotting each numerical variable against the target variable (See Appendix 5).

3. MODEL SELECTION

1. PERFORMING SINGULAR LINEAR REGRESSION MODELS

Our model selection process started with creating single variable regressions against IMDB scores for all variables. From these regression models the associated p-values, r_squared values and coefficient values were analyzed and stored. The storage of these values occurred within an external dataframe. (See Appendix 6)

2. NCV TESTS FOR HETEROSKEDASTICITY

The singular linear regression models, underwent a Non-Constant Variance (NCV) test. The

3. MODEL SELECTION

NCV test is used for addressing potential heteroskedasticity, a condition where the variance of errors varies across observations. The NCV test is used for addressing potential heteroskedasticity, a condition where the variance of errors varies across observations. The models that exhibited heteroskedasticity with statistical significance ($p<0.05$) had Robust standard errors used to adjust the estimates to eliminate heteroskedasticity .Think of a garden hose. When the water pressure is consistent (homoskedastic data), the flow is steady and predictable. However, if there are areas where the hose is pinched or widened (heteroskedastic data), the water flow becomes erratic and unpredictable. Applying robust standard errors is like installing a regulator that ensures an even flow of water throughout, regardless of the hose's varying conditions.. After performing this we now have a list of significant predictors without heteroskedasticity. (See Appendix 7 and Appendix 8).

3. DETERMINING TUKEY TEST VALUES FOR ALL VARIABLES

We conducted a Tukey Test on all predictors. The Tukey Test acts as a way to check if the regression models between the independent variable and dependent variable (IMDB score) perform better in a nonlinear function.Imagine baking cookies with a basic recipe. Someone suggests adding salt to enhance the flavor. You bake two batches: one with the usual recipe and one with added salt. The batch with salt tastes better, showing a small change can make a big difference. This is like adding a quadratic term to a linear model. The basic recipe is your original model, and the salt represents the quadratic term. Just like the salt improves the cookies, the quadratic term can enhance the model if it leads to better results, much like how the Tukey test evaluates the addition's impact. The adjustments to the regression models were only taken if they were statistically significant with a p value <.1. (See Appendix 9). Residual plots of the variables identified within the Tukey Test were used to further assess the non-linearity of the variable in relation to the dependent variable of IMDb Score. (See Appendix 10 Figures 1-11)

4. MULTICOLLINEARITY TEST

An additional step that was taken to further refine the variables within our model was a Multicollinearity test. This test is used to understand in which two or more predictor variables in a multiple regression model are highly correlated. Having multicollinearity can be

3. MODEL SELECTION

problematic because it makes it difficult to assess the individual contributions of a variable. This is the last and final analogy we promise, but imagine you are coaching the Canadian Olympic Rowing team and you are trying to determine which individuals who are contributing the most to the success of the team.The synchronization amongst the team members, similar to multicollinearity, blurs who contributes more to the boat's speed, making it impossible to make the assessment of who the better rower is. This is similar to the difficulties in assessing individual predictors' impacts in a model with multicollinearity. VIF is a tool that is used to assess how correlated variables are, typically we would have to remove a variable if the VIF within a relationship between variables is greater than 4 lucky for us this did not occur. (See Appendix 9)

5. SELECTION AND ASSIGNING DEGREE TO VARIABLES

The culmination of our model selection process involved several crucial steps: filtering predictors based on statistical significance and R-squared values, assessing multicollinearity, and subsequently constructing a linear regression model.

To determine the most relevant predictors for our model, we applied specific criteria. First, we selected variables that showcased a p-value of less than 0.05, indicating their statistical significance. Alongside this, we prioritized predictors with an R-squared value greater than 0.01, ensuring they individually explained a minimal variance in the outcome, yet collectively contributed to the model's accuracy. The cut-off for the R-squared values was determined using an elbow plot (see Appendix 13), which helped us identify the inflection point and thus refine our variable selection process. Refined predictor selection was used to build a Generalized Linear Model (GLM), and evaluate it using k-fold cross-validation. The non-linear relationship was further explored by creating models with combination of different degrees (polynomial terms), comparing them using ANOVA, and selecting the best-fitting model (see Appendix 12). Finally, to reassess the selected model's performance with cross-validation, Mean Squared Error was calculated (MSE).

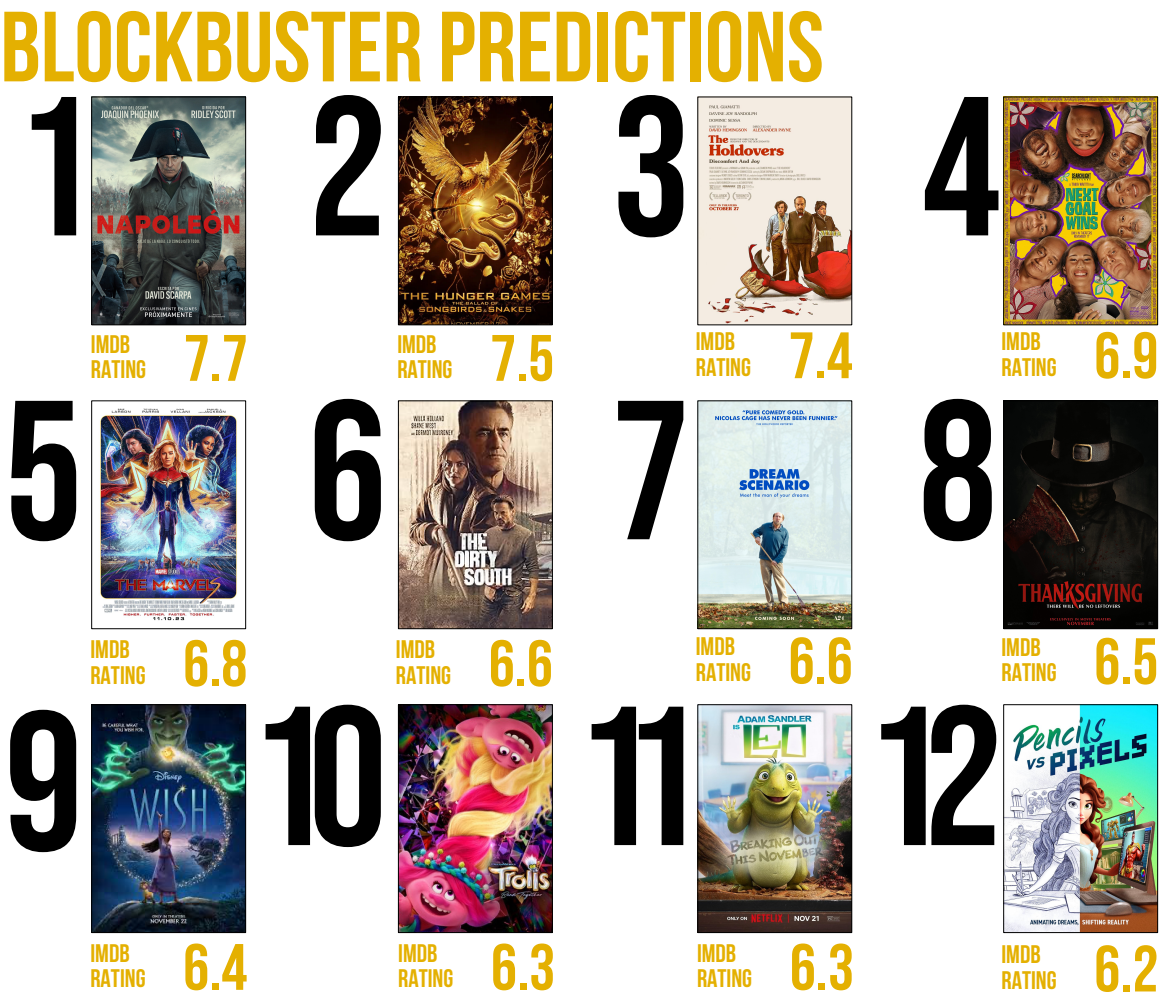
4. RESULTS

FINAL MODEL

Our strategy was to methodically identify the most relevant predictors by starting with dummying every categorical variable including the keywords. - this gave us 11302 variables. After our process we narrowed it down to 29. After all model selection steps the final model that has been selected is (See Appendix 14).

CONCLUSION

The model's performance was evaluated using the coefficient of determination (R-squared), which was found to be 0.423. This metric reflects the proportion of the variance in the dependent variable that is predictable from the independent variables. In terms of predictive power, the model's out-of-sample performance was assessed using k-fold cross-validation with



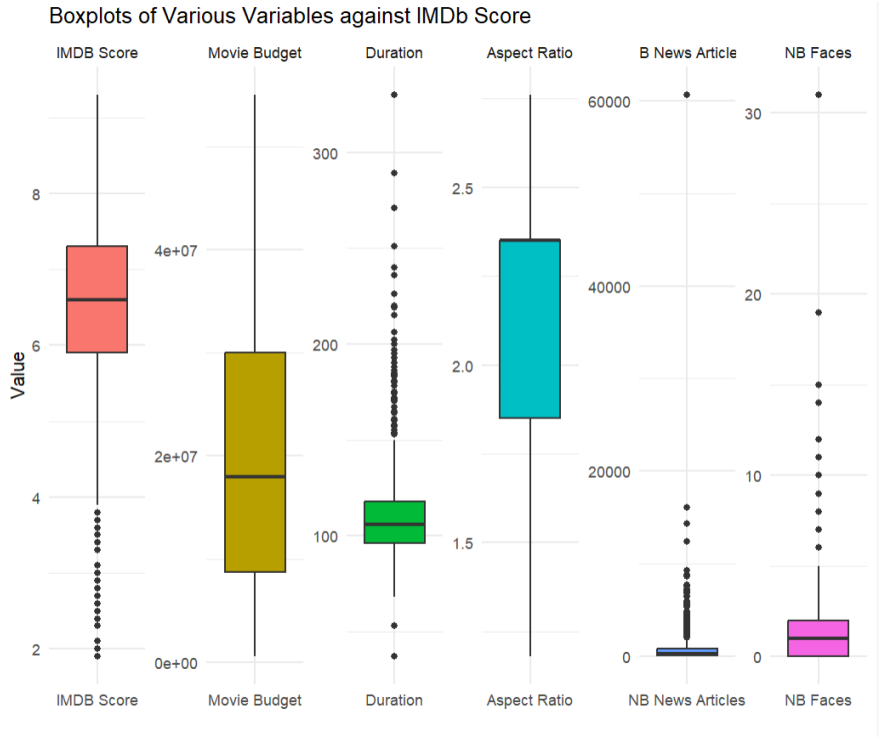
4. RESULTS

k=10. The mean squared error (MSE) from this cross-validation was 0.685986639021417, which provides an estimate of the model's predictive accuracy on new, unseen data. The significance of each predictor was evaluated based on p-values from the regression analysis, with a threshold of $p < 0.05$ for statistical significance. The predictors included a range of variables such as movie budget, release year, number of news articles, and star meter of actors, among others. For detailed statistical results, including the significance of each predictor, please refer to Appendix 14, which presents the final model details using the stargazer package.

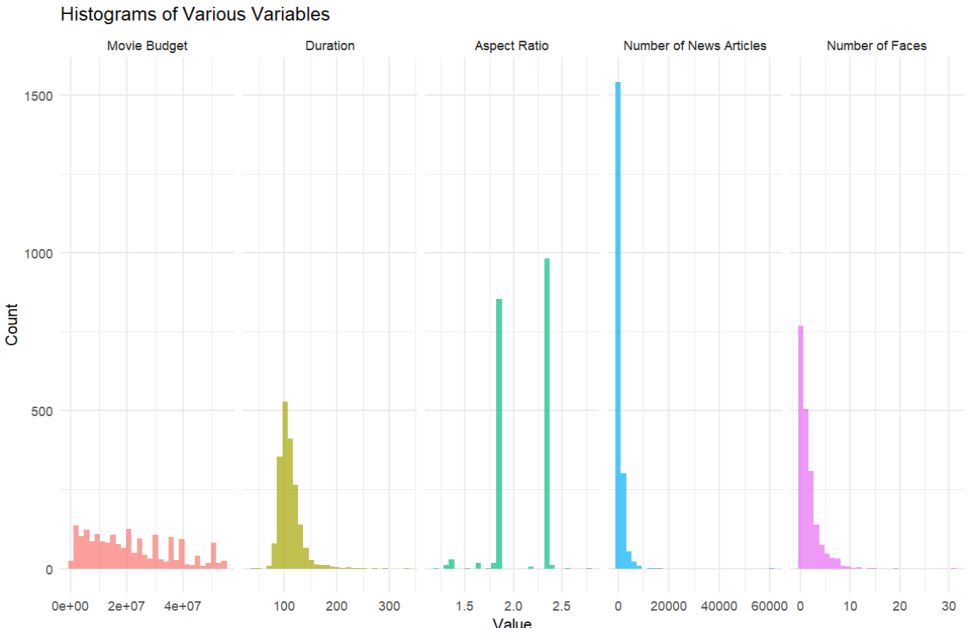
Furthermore, the model accounted for potential non-linear relationships and interaction effects between predictors, ensuring a robust approach to capturing the complexities of the data. Non-linear terms and polynomial expansions were considered for variables such as movie budget and release year, with the best-fitting terms selected based on the Akaike Information Criterion (AIC) within a stepwise regression framework. In conclusion, the final model demonstrates a solid ability to predict IMDb scores, with a strong R-squared value and satisfactory predictive power. The comprehensive statistical analysis, including the significance of predictors and model diagnostics, is documented in Appendix 14.

APPENDICES

APPENDIX 1: BOX PLOTS OF VARIOUS VARIABLES AGAINST IMDB SCORE



APPENDIX 2: HISTOGRAMS OF VARIOUS VARIABLES



APPENDICES

APPENDIX 3: TOP 10 POSITIVE CORRELATIONS

Top Positive Correlations

Variable	Correlation
Duration	0.39645091
Drama Genre	0.34556353
Number of News Articles	0.2483097
Biography Genre	0.1755708
Black and White Film Colour	0.15906912
Country of Origin - UK	0.12570925
R Rating	0.12441561
History Genre	0.10485668
Approved Rating	0.10124239
Distributed from the USA	0.09153258

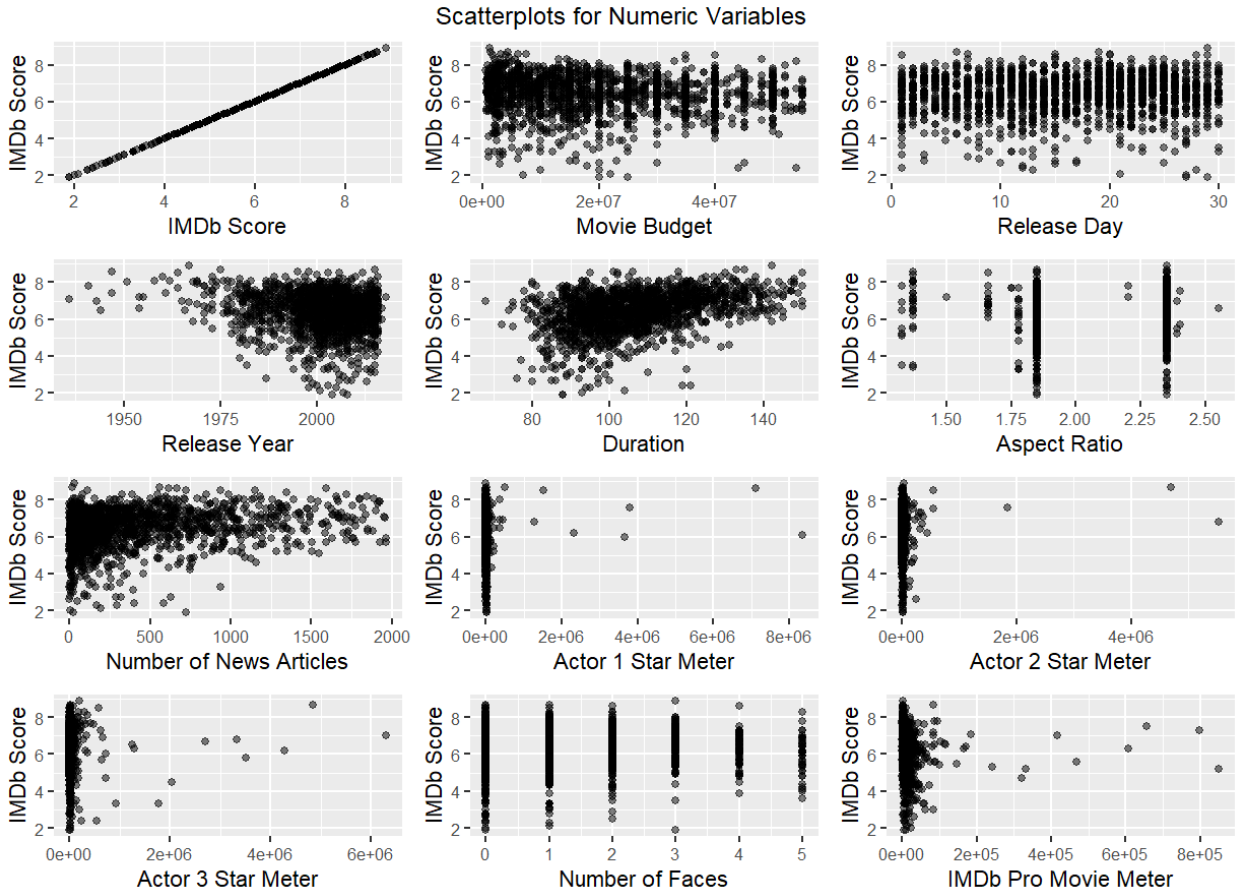
APPENDIX 4: TOP 10 NEGATIVE CORRELATORS

Top Negative Correlations

Variable	Correlation
Horror Genre	-0.1929524
Release Year	-0.1639177
Film Colour	-0.1590691
Action Genre	-0.1542171
Comedy Genre	-0.1501395
PG 13 Rating	-0.1479250
Cinematographer-Shawn Maurer	-0.1470642
Director-Jason Friedberg	-0.1372913
Actor-Carmen Electra	-0.1366294
English	-0.1248566

APPENDICES

APPENDIX 5: SCATTER PLOTS FOR NUMERIC VARIABLES AGAINST IMDB SCORE

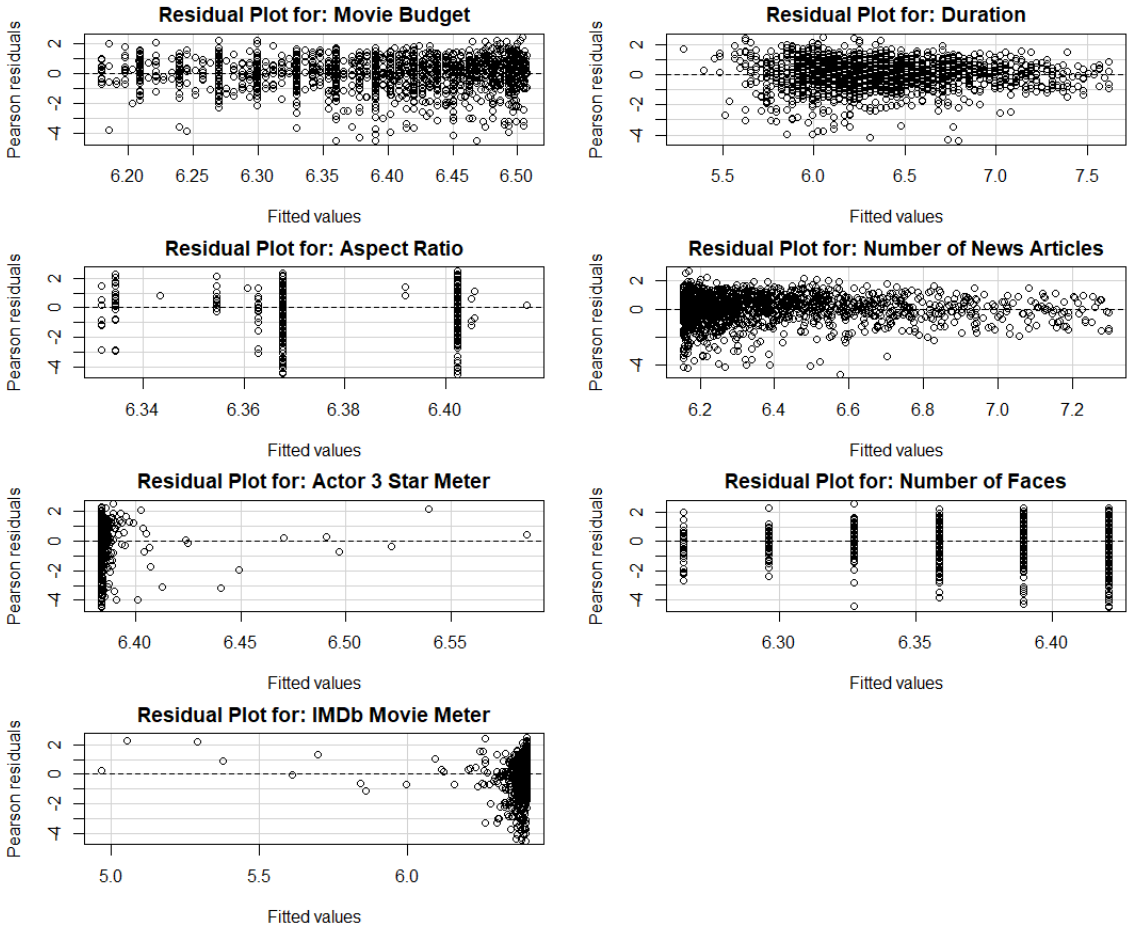


APPENDIX 6: DATA FRAME STORING COEFFICIENTS, R-SQUARED AND P-VALUES

	Variable	Coefficient	P_Value	R_Squared
movie_budget	movie_budget	-6.04E-09	0.000753654	0.00710753
release_day	release_day	0.004361445	0.169538702	0.00118498
release_year	release_year	-0.015607231	4.58E-11	0.02686899
duration	duration	0.028511755	3.88E-61	0.15717332
aspect_ratio	aspect_ratio	0.069370413	0.477739035	0.00031664
nb_news_articles	nb_news_articles	0.000582251	2.27E-23	0.06043286
actor1_star_meter	actor1_star_meter	1.34E-07	0.111964113	0.00158609
actor2_star_meter	actor2_star_meter	2.84E-07	0.041100667	0.00261790
actor3_star_meter	actor3_star_meter	3.22E-08	0.731326763	7.41E-0
nb_faces	nb_faces	-0.03119945	0.127121096	0.00146127
movie_meter_IMDBpro	movie_meter_IMDBpro	-1.70E-06	0.004919205	0.00495713
genresDrama	genresDrama	0.730610072	6.29E-46	0.11941415
genresBiography	genresBiography	0.739796617	1.68E-12	0.03082028
genresSport	genresSport	0.329151523	0.006816404	0.00458932
genresHorror	genresHorror	-0.630841228	7.81E-15	0.03723064
genresThriller	genresThriller	-0.143201908	0.011682876	0.00398723

APPENDICES

APPENDIX 7: RESIDUAL PLOTS FOR HETEROSKEDASTICITY



APPENDIX 8: LIST OF VARIABLES PRESENTING HETEROSKEDASTICITY

Variables Presenting Heteroskedasticity

Variable	P-value
Movie Budget	1.734919e-06
Release Day	0.8822285
Release Year	0.4665053
Duration	4.706085e-21
Aspect Ratio	5.629506e-06
Number of News Articles	0.0002515216
Actor 1 Star Meter	0.5893437
Actor 2 Star Meter	0.6297132
Actor 3 Star Meter	0.01356959
Number of Faces	8.349851e-05
IMDBpro Movie Meter	0.004970374

APPENDICES

APPENDIX 9: NON-LINEARITY TUKEY TEST

Tukey Test Analysis Results

Variable	Test Stat	Pr(>|Test Stat|)	Significance
Movie Budget	1.8616	0.06284	•
Release Day	-0.6231	0.5333	
Release Year	2.4666	0.01374	*
Duration	-0.0922	0.9266	
Aspect Ratio	1.5892	0.1122	
Number of News Articles	-3.8555	0.0001201	***
Actor 1 Star Meter	-1.1366	0.2559	
Actor 2 Star Meter	-1.0448	0.2963	
Actor 3 Star Meter	1.8611	0.06291	•
Number of Faces	-1.7006	0.08922	•
Movie Meter IMDb Pro	6.1856	7.848e-10	***

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

Singnificantly Non-Linear Variables:

- Movie Budget
- Release Year
- Number of News Articles
- Actor 3 Star Meter
- Number of Faces
- Movie Meter IMDb Pro

APPENDIX 10: RESIDUAL PLOTS OF SIGNIFICANTLY NON-LINEAR VARIABLES

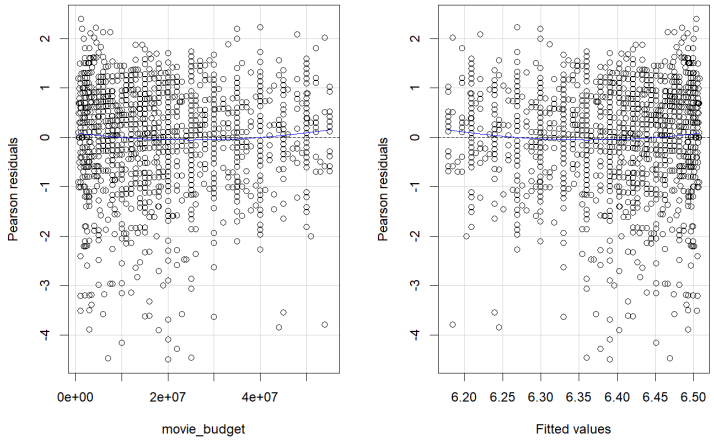


Figure 1: Movie Budget

APPENDICES

APPENDIX 10: RESIDUAL PLOTS OF SIGNIFICANTLY NON-LINEAR VARIABLES

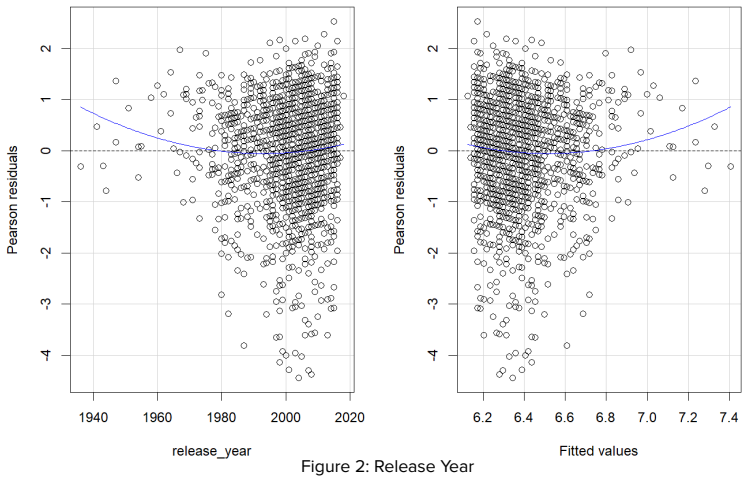


Figure 2: Release Year

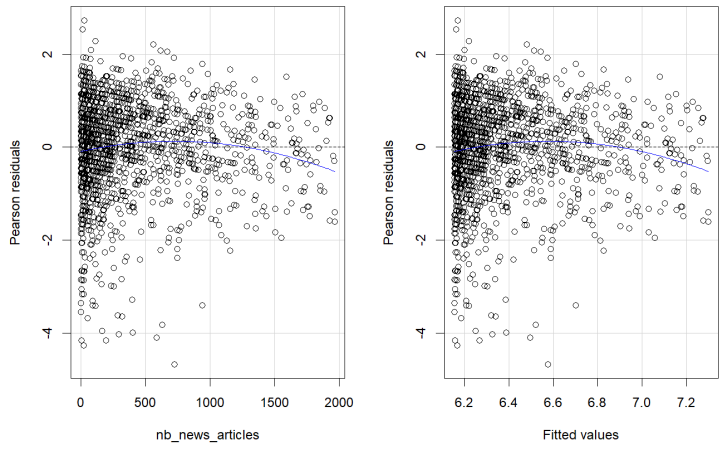


Figure 3: Number of News Articles

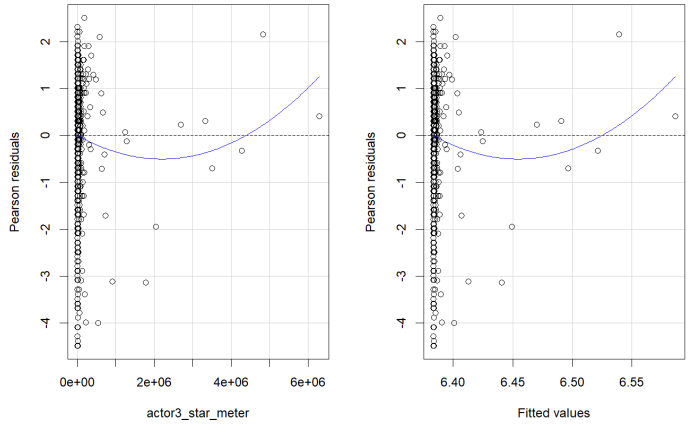


Figure 4: Actor 3 Star Meter

APPENDICES

APPENDIX 10: RESIDUAL PLOTS OF SIGNIFICANTLY NON-LINEAR VARIABLES

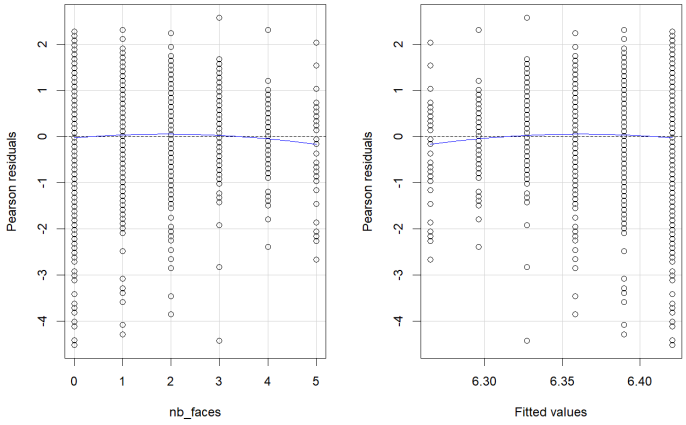


Figure 5: Number of Faces

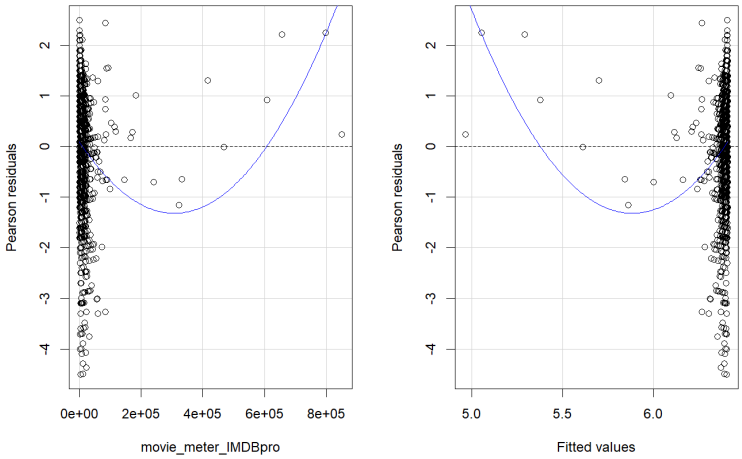


Figure 6: IMDbPro MovieMeter

Appendix 11: VIF Scores

Variable	Value
Movie Budget	1.115118
Release Year	1.108415
Duration	1.117220
Number of News Articles	1.098312
Actor 2 Star Meter	1.011654
IMDBpro Movie Meter	1.047252

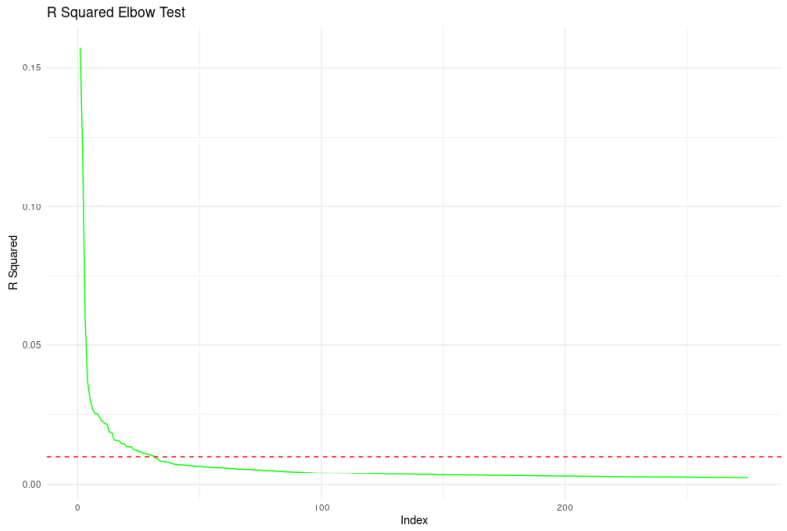
APPENDICES

APPENDIX 12: ANOVA TEST RESULTS

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)	
1	1564	1035.2					
2	1563	1024.3	1	10.8775	16.5915	4.869e-05	***
3	1562	1022.3	1	2.0077	3.0624	0.0803198	.
4	1561	1021.7	1	0.5898	0.8996	0.3430340	
5	1560	1021.6	1	0.1422	0.2169	0.6414723	
6	1563	1035.1	-3	-13.5032	6.8655	0.0001358	***
7	1562	1024.0	1	11.0638	16.8757	4.199e-05	***
8	1561	1022.1	1	1.9586	2.9875	0.0841059	.
9	1560	1021.5	1	0.5298	0.8081	0.3688198	
10	1559	1021.4	1	0.1523	0.2323	0.6299054	
11	1562	1033.3	-3	-11.8840	6.0423	0.0004341	***
12	1561	1022.9	1	10.3783	15.8301	7.248e-05	***
13	1560	1021.1	1	1.8285	2.7890	0.0951119	.
14	1559	1020.5	1	0.5764	0.8792	0.3485782	
15	1558	1020.4	1	0.1331	0.2030	0.6523638	
16	1561	1033.3	-3	-12.9161	6.5670	0.0002070	***
17	1560	1022.9	1	10.3836	15.8381	7.217e-05	***
18	1559	1021.1	1	1.8248	2.7833	0.0954505	.
19	1558	1020.5	1	0.5753	0.8775	0.3490459	
20	1557	1020.4	1	0.1329	0.2027	0.6526314	
21	1560	1033.1	-3	-12.7566	6.4859	0.0002322	***
22	1559	1022.6	1	10.4792	15.9840	6.687e-05	***
23	1558	1020.8	1	1.8195	2.7753	0.0959275	.
24	1557	1020.2	1	0.5702	0.8698	0.3511675	
25	1556	1020.1	1	0.1289	0.1966	0.6575723	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

APPENDIX 13: R-SQUARE ELBOW TEST RESULTS



APPENDICES

APPENDIX 14: STARGAZER TABLE

Final Model Results

	Dependent variable:
	IMDB Score
Duration	0.013*
	(0.002)
Genres-Drama	0.323*
	(0.051)
Genres-Biography	0.224**
	(0.091)
Genres-Horror	-0.492*
	(0.073)
Genres-Comedy	-0.113**
	(0.049)
Genres-Sci-Fi	-0.081
	(0.070)
Genres-Action	-0.350*
	(0.056)
Genres-History	-0.039
	(0.134)
Plot Keywords-Punctuation in Title	-1.876*
	(0.481)
Plot Keywords-Box Office Flop	-0.567**
	(0.247)
Plot Keywords-Critically Bashed	-0.582**
	(0.288)
Plot Keywords-Masturbation	-1.367*
	(0.366)
Plot Keywords-Mousetrap	-1.409**
	(0.692)
Plot Keywords-Evil	-1.132*
	(0.335)
Language-English	-0.700*
	(0.151)
Country-UK	0.108
	(0.097)
Country-USA	-0.131*
	(0.072)
Maturity Rating-Approved	0.126
	(0.276)
Maturity Rating-PG 13	-0.014
	(0.067)
Maturity Rating-R	0.232*
	(0.063)
Director-Don Michael Paul	-2.284*
	(0.576)
Director-Jason Friedberg	-0.184
	(1.164)
Director-Uwe Boll	-2.800*
	(0.578)
Actor1-Carmen Electra	-0.960
	(0.811)
Colour Film-Black and White	0.456*
	(0.126)
Cinematographer-Shawn Maurer	-1.319*
	(0.509)
Production Company-New Regency Pictures	-0.780
	(0.587)
Release Year	-8.813*
	(1.000)
Release Year^2	-0.721
	(1.007)
Release Year^3	1.129
	(0.858)
Nb News Articles	10.601*
	(0.876)
Nb News Articles^2	-3.266*
	(0.821)
Constant	5.710*
	(0.231)
Observations	1,594
R^2	0.423
Adjusted R^2	0.411
Residual Std. Error	0.809 (df = 1561)
F Statistic	35.749* (df = 32; 1561)
Note:	*p<0.1; **p<0.05; ***p<0.01