McGill University

Multivariate Statistics

MGSC 661

Midterm Project

Louis D'hulst (260768615)
Hadyan Fahreza (260986028)
Shivanshu Gupta (260986028)
Nadine Hamra (260666146)
Rebecca Mukena Yumba (261003552)

January 27, 2022


R^2D2
R
RESTRICTED
AGES UNDER TWENTY JUAN REQUIRES
ACCOMPANYING PARENT OR ADULT GUARDIAN
®

1 Introduction

With this project, we have been given the opportunity to understand the driving forces behind the success of great films. Our team was given a dataset containing the IMDB ratings of over 2000 movies along with relevant information such as budget and year of release. The goal of the project is to build a model to predict the IMDB score of a movie based on significant predictive variables in the dataset. The model should be able to accurately predict the score of movies that vary across several dimensions such as language, duration, genre, year of release, production country, and budget.

While an IMDB score cannot perfectly represent the subjective value of how good or enjoyable a movie is, it is safe to assume that quality movies are rated more highly on average. Ratings are frequently used to decide if a movie is worth watching or not, so a highly rated movie is likely to attract more viewers than a lower rated one. Our model would therefore be able to tell producers and production companies what they should focus on in order to create a highly rated movie. It is important to note that we do not claim to make a model that can tell producers how to make a good movie; the movie-making process is far too complicated for that.

The purpose of the project is to implement concepts learned from the class and to gain actual hands-on experience of creating a full-fledged regression model. Firstly, we will parse through and analyze the data to get an idea of what variables would be considered relevant and which should be discarded. The outcome will be a modified data set for us to use to help build our model. Next, we will run through variations of splines, regressions, and cross validation tests to develop a predictive model with a lowest optimal MSE value. For those who aren't familiar with statistics, a lower MSE (mean squared error) is indicative a powerful predictive ability. Finally, we will discuss the managerial implications of our findings and determine the impact of our predictors on the potential outcome of an IMDB rating.

2 Data Description

2.1 Distributions

We began by investigating variables individually to study their distributions through histograms (Figure 2 in Appendix) and box plots. The two plots provided us with an idea of their skewness, range, and outliers. This information would help down the line during model selection.

For example, the year_of_release histogram shows that most observations are clustered around the mid-2000s and a major exponential increase around the 1980s. Since there are more observations around the mid-2000s, the model is less sensitive to predicting IMDB scores for movies in the mid-2000s relative to those in the 80s simply because there are more observations in the mid-2000s. The scatter plot validates this statement: there is more variation of IMDB scores in the mid-2000s; as such, the model will be less sensitive to a highly variant movie released in the mid-2000s than one released in the 40s or 50s, for example.

2.2 Scatter Plots

To have a better understanding on how each predictor affects the target variable (imdb_score), we plotted each predictor against the target variable. We are more interested in the scatter plots of the continuous predictors. Since the values of the categorical variables and dummy variables are discrete (0 or 1 for dummies), their scatter plots will be not as relevant as the continuous ones. From the scatter plots in Figure 3 of the Appendix, we can see that the interaction between the continuous predictors and the target variable are all non-linear. Thus, when working on the regression model, we want to make sure that our model takes into account the interactions, either in the form of transformations or polynomial interactions.

2.3 Collinearity

Correlation and collinearity are evaluated using the correlation matrix function, cor(), in R. We followed the general rule of thumb: if the correlation between two predictors is more than 0.8, they are considered collinear variables. From the collinearity matrix in Figure 4 of the Appendix, it can be observed that the highest correlation value across the matrix is 0.4375, meaning we are safe to use these variables without worrying about collinearity.

2.4 Outliers

To find and eliminate outliers, we ran outlier tests on simple linear regression for each predictor separately. We found and removed 2 observations that stood out consistently in most of our outlier tests: observations 633 and 895. We also observed an additional outlier in the boxplot for the total number of actors that did not appear in the outlier test. The observation is where the total number of actors is equal to 313, and can be seen as an outlier in Figure 1 below.

After removing the outliers from the dataset, the team ran a regression model and compared the p-value/R-square of the new model with the p-value/R-square of the regression with outliers. It appeared the significance of the predictors and the multiple R-square for the whole model was significantly improved.

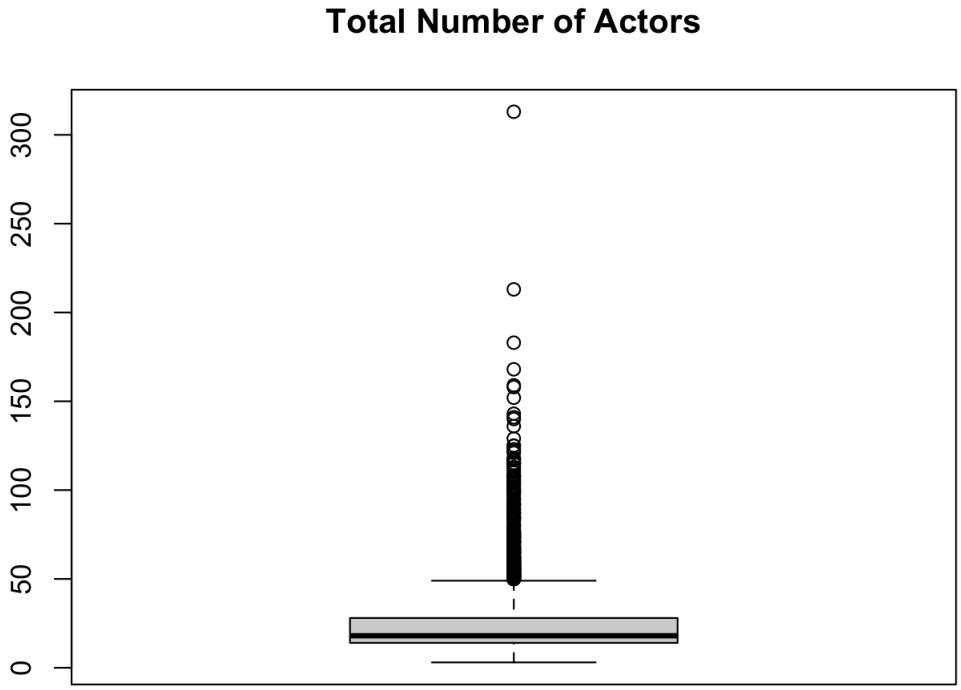

Figure 1: Log transformation of budget_in_millions

2.5 Heteroskedasticity

Heteroskedasticity happens when the variance of the errors of the regression function are non-constant along the horizontal axis (increasing or decreasing). It poses a problem since the standard errors produced by the

heteroskedastic predictors are biased. To detect heteroskedasticity, the team did both visual test on the residuals scatter plot of each predictors and the non-constant variance (NCV) test for the predictors flagged from the visual test. See figure 5.

From the visual test, it can be seen that the following predictors have the possibility of being heteroskedastic:

1. budget_in_millions
2. year_of_release
3. duration_in_hours
4. total_number_of_languages
5. total_number_of_actors
6. total_number_of_directors
7. total_number_of_producers
8. total_number_of_production_companies
9. total_number_of_production_countries

NCV tests on each of the flagged predictors were run to verify the suspicion. See Table 3. If the NCV test yields p-value less than 0.05, the predictor is heteroskedastic. The residual plots in Figure 5 of the Appendix gives some indication as to which predictors suffered from heteroskedasticity.

The NCV test determined the following predictors to be heteroskedastic.

1. year_of_release
2. duration_in_hours
3. total_number_of_actors
4. total_number_of_producers

2.5.1 Standardizing Heteroskedasticity

Heteroskedasticity can be corrected for by using the robust standard error formulation in the linear regression model. This formulation provides standard errors that are capable of robust inference even in the presence of heteroskedasticity. The p-values of a model created using robust standard errors are therefore valid. Heteroskedasticity does not affect coefficients, however, meaning that predictions made using a model that does not utilize robust standard errors are still valid. The final prediction model does not need to account for heteroskedasticity.

2.6 Simple Linear Regressions

We wanted to understand the relationships of individual predictors with the target variable. Specifically, we wanted to determine which individual predictors had the strongest relationships to imdb_score and the magnitude of their impact on the target variable. It is important to check for nonlinearity to have a better understanding of how to treat each variable relative to its relationship with the target variable. We will dive further into the findings of the linear regression below.

2.6.1 Continuous Variables

We ran linear regressions between the target variable IMDB_score and all the quantitative variables. The goal was to determine which of these variables most significantly affected the IMDB score based on the p-value. Using 0.05 as our threshold, it was determined that the total number of directors, total number of production companies, and total number of production countries should be excluded from our model since they had a statistically insignificant relationship with the target variable. Note that since some predictors were determined to suffer from heteroskedasticity, we computed robust p-values for all predictors to get accurate results. Table 1 of the appendix shows the coefficients, p-values and R^2 of the simple linear regressions. Table 2 shows the coefficients and robust p-values of the simple linear regressions.

2.6.2 Dummy Variables

We took a similar approach with dummy variables. We used the same function to run linear regressions between the target variable, imdb_score, and the dummy variables available to us in the dataset . Once again, we used the p-value as a reference to identify which variables should be considered in our model and all dummies with a p-value greater than 0.05 were discarded. We were left with: genre_action, genre_animation, genre_comedy, genre_biography, genre_crime, genre_drama, genre_family, genre_fantasy, genre_music, genre_musical, genre_romance, genre_scifi, genre_sport, genre_filmnoir, genre_horror, genre_thriller, main_actor1_is_female, main_actor2_is_female, and main_actor3_is_female. Table 1 of the appendix shows the coefficients, p-values and R^2 of the simple linear regressions. Table 2 shows the coefficients and robust p-values of the simple linear regressions.

2.6.3 Categorical Variables

To draw meaningful interpretations of the categorical variables, we specified the variable type to be factor (categorical, enumerative) rather than numeric using the as.factor() function in R. We then ran simple linear regressions and observed the r-squared values, since it would be difficult to assess the attributes individually. This gave us an idea of which categories would be useful in interpreting the IMDB score.

2.7 Model Selection

There are several steps involved in model selection. Based on the above analysis done on all the predictors, we will walk through our rationale behind developing our benchmark model and the course of action leading up to our final predictive model.

The variables are classified into quantitative, categorical, and dummy variables, as such, we went through each category to determine which variables within them should be included. The principles behind how we selected the specific predictors is discussed in the following section.

2.7.1 Benchmark Model and Polynomial Regressions

As a foundation, we first started with the significant quantitative variables that we had identified as being significant in our simple linear regressions. We ran multiple polynomial regressions between each selected variable and the imdb_score. See figure 6. In order to find the best fit, we used the ANOVA test to evaluate which degree to use for each predictor individually. ANOVA indicated whether increasing the degree of the polynomial resulted in a significant improvement in the relationship of a predictor and the target variable. Based on the significance we chose the best degree for each predictor.

Based on results from the ANOVA tests, we settled on the model shown in Table 4 of the Appendix. This model gave an MSE of around 0.673.

2.7.2 Cross Validation

Then, we ran a nested for-loop cross validation k-fold test, varying the degrees of the polynomials for each predictor in order to find the optimal combination that will minimize the mean squared error (MSE). The optimal solution gave an MSE of around 0.659.

2.7.3 Splines

Looking at the data points for the duration_in_hours scatter plot, there seems to be an initial upward trend that flattens out after the 2 hour point. A spline was added at 2.2 hours to capture this difference in trend. Polynomials of degree 1 through 5 were fitted. Adding the spline and polynomials to the scatter plot confirms the hypothesis, as seen in Figure 7 of the Appendix. The shape of the cubic and quartic splines are similar to that of the regular quintic polynomial in Figure 6. The R^2 of the quartic spline is 0.1495 compared to 0.1492 for the non-spline quintic polynomial, justifying the use of a spline.

Visual analysis of the scatter plots led us to consider adding splines for two predictors: year_of_release and duration_in_hours. The data points for the year_of_release plot start in a fairly linear pattern then grow into a funnel shape as the years increase. A spline was added at year 1970 to try and capture the increase in movies with a rating below 7. Polynomials of degrees 1 through 5 were fitted. Interestingly, Figure 8 shows a steeper decrease before the spline and a flatter decrease after, indicating that the large number of movies with ratings below 7 is more than offset by the number of movies with higher ratings. While the shape of the spline polynomials is similar to that of regular quadratic regressions (Figure 6), the R^2 of the quadratic spline is 0.0912 while that of the quadratic regression is 0.07827, justifying the use of a spline for year_release in the final model.

Adding these spline to the model gave an MSE of around 0.655.

2.7.4 Adding Dummy Variables

We then looked into adding dummy variables. All of the dummies in the dataset were put into a multiple linear regression. We then filtered out dummies that were not significant below the 5% level, and further filtered out those that were not significant below the 1% level., giving us two lists of dummies. The dummies in the latter list were added to the predictive model first, giving an MSE of around 0.585. The dummies in the former list were then added to the predictive model, giving an MSE of around 0.574; the idea was to verify if adding so many dummies led to overfitting. The results show that having the dummies significant at the 5% level led to better out-of-sample performance, so we kept them in the model.

2.7.5 Adding Categorical Variables

For the final grouping, the R-squared of each categorical variable was used to determine if it will be included in the final model or not. Since it would be difficult to segment them based on individual p-values, we opted for variables with high R-squared values, using anything above 0.5. Actors, directors, producers,and production companies were the only variables with R-squared values greater than the threshold.

To determine the impact of highly-performing actors, directors, producers, and production companies on the IMDB score of their movies, we generated dummy variables to indicate whether the actor, director, producer, or production company had a higher than average IMDB score (1) or not (0). However, when calculating the average IMDB score, we excluded the IMDB title in the corresponding row. We did this to not include the effect of the movie itself in predicting its IMDB score. We created a correlation matrix of these new predictors to see if there would be any issue with collinearity. Figure 11 of the Appendix shows that the highest absolute value for the correlations is 0.28, so collinearity is not a problem.

Adding these newly computed dummy variables to the predictive model led to an MSE of around 0.538, thereby justifying their inclusion in the model.

2.7.6 Interaction Terms

Visual analysis of scatter plots led us to an interaction effect. Indeed, as the year increases the budget of the movies increases rapidly, as shown by Figure 9 of the Appendix. We tried adding this interaction effect to the predictive model, giving an MSE of 0.537. Since the MSE was lowered, the addition is justified.

2.7.7 Cleaning the Model

The predictive model now has many variables in it. To check if we were overfitting by including so many predictors, we tried removing some of them. Since year_of_release had been added in the interaction term, we began by removing the polynomial year_of_release predictor. This lowered the MSE to 0.535, justifying the removal. We then tried removing several other predictors, none of which made the model perform better except for month_of_release. Removing month_of_release lowered the MSE to 0.533.

2.7.8 Predictor Transformation

In an attempt to lower MSE even further, we thought of transforming some of the predictors. The histogram for budget_in_millions (Figure 10) shows an inverse exponential shape. Log transforming this predictor makes it more bell-shaped.

Including this log transformation in the model brought the MSE down to 0.528. The scatter plot for the interaction term between budget_in_millions and year_of_release has an exponential shape. Log transforming this interaction term yielded an MSE of around 0.527.

2.7.9 Final Polynomial Optimization

Given that the model has changed substantially since the last polynomial degree optimization, we optimized another time in order to get the best possible mean-squared error. The average MSE from this optimization was 0.5268.

3 Managerial Interpretation

It is evident that there are many factors that can affect the quality and rating of a movie. From the month the movie is released down to the editor's name, we sorted through a wide range of data to pinpoint the most influential factors movie producers should consider.

As a disclaimer, it is difficult to interpret the impact of a polynomial expression through their coefficients. We will address them by discussing their influence based on their behaviour outside our model and providing a general overview of how they influence the IMDB score.

Talking about the regression model, the decision to use polynomial regression instead of linear regression is to give the model the ability to freely fit to the data points used in the model. From the scatterplot (Figure 3) we generated during the statistical analysis process, it can be seen that most of the predictors are not positioned in a linear fashion. Interactions of higher degree are needed to address the non-linear distribution of the predictors. Throughout the series of statistical analyses we have done, the team were able to extract some of the most important factors from the dataset. This is done to reduce the amount of time needed to develop the model. Also, there is an overfitting issue related to taking into account all the predictors to the model. It is possible by doing it, it will hurt the prediction model when predicting the IMDB score of a movie that is not in the dataset we used for developing the model. The model will try to fit to all data points used to generate the

model as close as possible, in which the slightest difference in the attributes of the prediction inputs can result in large shifts in IMDB score prediction.

Based on the statistical analysis, it is known that the essential factors determining the success of a movie are production budget, year of release, movie duration, number of actors and producers involved in it, month of release. In addition to that, the types of genre, along with whether the main actor is female, also have an impact in ruling on the success of a movie.

First we can focus on the characteristics of the movie. By knowing the genre of the movie, specifically looking at the ones listed in the model, it would give you an idea of what the score could be. Next would be the duration of the movie. According to our analysis,increasing the length of the movie tends to increase the IMDB score.

Renowned Hollywood names obviously play a big part in the final movie rating. Whoever is responsible for hiring the hierarchy of a movie crew should prioritize the actors, directors, and producers. Factors such as editor name or main languages spoken will not play a big part in helping determine the movie's rating. If the audience sees Pixar in the movie intro, they immediately know to prepare themselves for an emotional roller coaster. The production company name is indeed another variable affecting the rating.

Then we can move onto the miscellaneous features. Surprisingly, a bigger budget does not guarantee a high IMDB rating. Mad Max had a higher rating than Avatar, even though the latter had more than a hundred million dollars available to them. It would be fair to infer that any movie termed 'classic' was likely released over twenty years ago. While producers don't have much control over the year of release, it would be interesting to note that the later the year you'll likely have a lower IMDB score. The month ,on the other hand, increases the score as we cycle through the months. The total number of actors or producers minimally affects the score and should not be a big concern for whosoever is creating the movie.

4 Appendix

Table 1: Simple linear regression results, ordered in ascending by p-value

	predictor	coefficient	pval	rsq
1	duration_in_hours	0.98	0	0.13
2	genre_drama	0.58	0	0.09
3	year_of_release	−0.02	0	0.08
4	genre_horror	−0.69	0	0.05
5	total_number_of_actors	0.01	0	0.05
6	genre_comedy	−0.39	0	0.04
7	genre_action	−0.33	0	0.02
8	genre_biography	0.63	0	0.02
9	month_of_release	0.04	0	0.02
10	genre_history	0.61	0	0.02
11	genre_filmnoir	1.35	0	0.01
12	main_actor1_is_female	−0.21	0	0.01
13	genre_war	0.53	0	0.01
14	genre_scifi	−0.27	0	0.01
15	total_number_languages	0.08	0.0000	0.01
16	genre_family	−0.33	0.0000	0.01
17	genre_fantasy	−0.27	0.0000	0.01
18	budget_in_millions	−0.002	0.0003	0.004
19	genre_documentary	0.85	0.002	0.003
20	genre_musical	0.43	0.002	0.003
21	genre_western	0.47	0.003	0.003
22	main_actor2_is_female	−0.08	0.02	0.002
23	main_actor3_is_female	−0.07	0.04	0.001
24	genre_crime	0.09	0.05	0.001
25	genre_thriller	−0.08	0.07	0.001
26	total_number_of_directors	0.09	0.09	0.001
27	genre_sport	0.17	0.13	0.001
28	genre_adventure	−0.06	0.15	0.001
29	total_number_of_production_companies	−0.01	0.16	0.001
30	genre_animation	0.10	0.25	0.0004
31	main_director_is_female	−0.07	0.46	0.0002
32	genre_mystery	0.04	0.48	0.0002
33	genre_music	0.07	0.53	0.0001
34	genre_romance	0.03	0.57	0.0001
35	total_number_of_production_countries	0.01	0.80	0.0000
36	total_number_of_producers	0.002	0.89	0.0000

Table 2: Simple linear regression results with robust p-values, ordered in ascending by p-value

	Predictor	Coefficient	Robust p-value
1	month_of_release	0.04	0
2	year_of_release	−0.02	0
3	duration_in_hours	0.98	0
4	total_number_languages	0.08	0
5	genre_action	−0.33	0
6	genre_biography	0.63	0
7	genre_comedy	−0.39	0
8	genre_documentary	0.85	0
9	genre_drama	0.58	0
10	genre_filmnoir	1.35	0
11	genre_history	0.61	0
12	genre_horror	−0.69	0
13	genre_war	0.53	0
14	main_actor1_is_female	−0.21	0
15	total_number_of_actors	0.01	0
16	genre_family	−0.33	0.0000
17	genre_musical	0.43	0.0000
18	genre_scifi	−0.27	0.0000
19	genre_fantasy	−0.27	0.0001
20	budget_in_millions	−0.002	0.0004
21	genre_western	0.47	0.01
22	main_actor2_is_female	−0.08	0.02
23	genre_crime	0.09	0.04
24	total_number_of_directors	0.09	0.04
25	main_actor3_is_female	−0.07	0.05
26	genre_thriller	−0.08	0.06
27	genre_sport	0.17	0.09
28	genre_adventure	−0.06	0.15
29	genre_animation	0.10	0.15
30	total_number_of_production_companies	−0.01	0.18
31	main_director_is_female	−0.07	0.41
32	genre_mystery	0.04	0.46
33	genre_music	0.07	0.50
34	genre_romance	0.03	0.54
35	total_number_of_production_countries	0.01	0.81
36	total_number_of_producers	0.002	0.88

Table 3: NCV test results

	Predictor	NCV p-value
1	budget_in_millions	0.586781219085894
2	year_of_release	1.91897059827664e-06
3	duration_in_hours	4.83347610364331e-10
4	total_number_of_actors	0.00012708938584331
5	total_number_of_producers	0.0143957076289458
6	total_number_of_production_companies	0.739396392472809

Table 4: Initial Model - Coefficients and Standard Errors

	Dependent variable:
	imdb_score
budget_in_millions	-8.679^{***}
	(0.973)
budget_in_millionsˆ2	9.158^{***}
	(0.864)
budget_in_millionsˆ3	-6.470^{***}
	(0.853)
budget_in_millionsˆ4	2.580^{***}
	(0.844)
budget_in_millionsˆ5	-1.427^{*}
	(0.835)
year_of_release	-6.639^{***}
	(1.055)
year_of_releaseˆ2	2.970^{***}
	(0.864)
duration_in_hours	17.420^{***}
	(0.898)
duration_in_hoursˆ2	-5.289^{***}
	(0.828)
total_number_of_actors	7.559^{***}
	(0.882)
total_number_of_actorsˆ2	-4.180^{***}
	(0.830)
total_number_of_producers	2.924^{***}
	(0.877)
total_number_of_producersˆ2	-3.155^{***}
	(0.825)
total_number_of_producersˆ3	2.108^{**}
	(0.830)
total_number_of_producersˆ4	-0.313
	(0.836)
total_number_of_producersˆ5	-0.368
	(0.826)
month_of_release	3.837^{***}
	(0.829)
month_of_releaseˆ2	0.693
	(0.827)
month_of_releaseˆ3	0.030
	(0.836)
Constant	6.711^{***}
	(0.015)
Observations	2,950
Log Likelihood	−3,571.338
Akaike Inf. Crit.	7,182.677
Note:	*p<0.1; **p<0.05; ***p<0.01

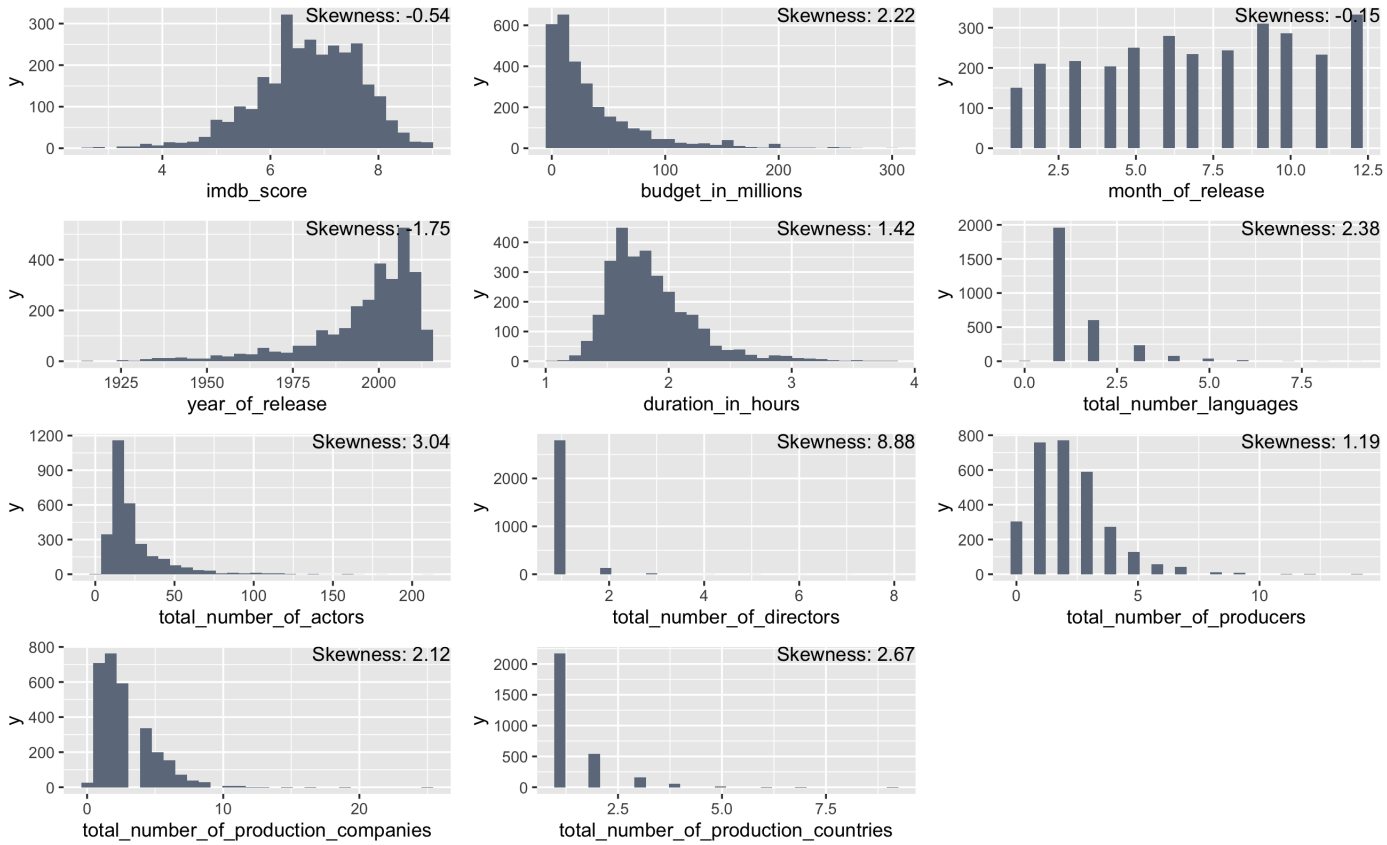

Figure 2: Histogram grid

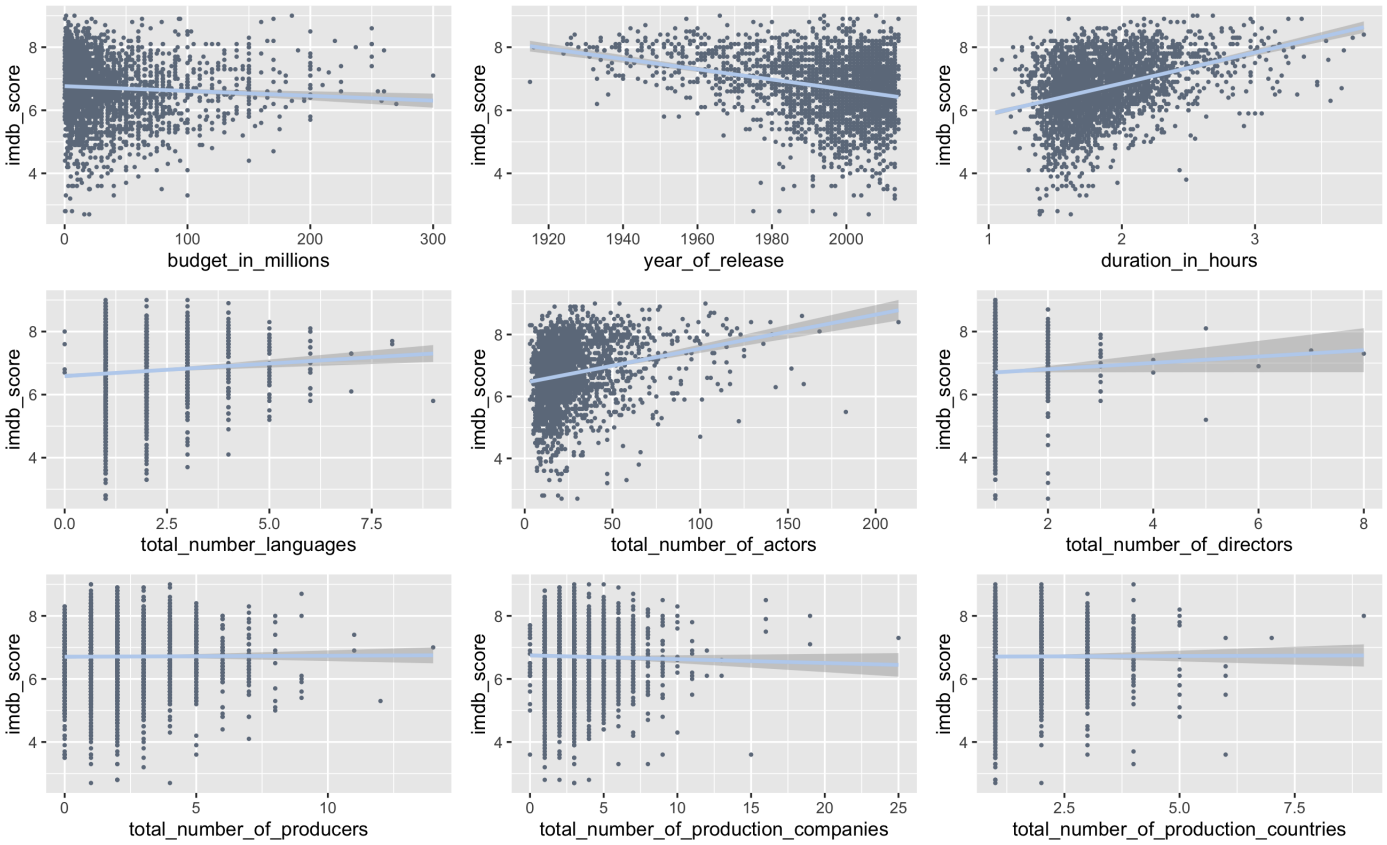

Figure 3: Scatter plot grid

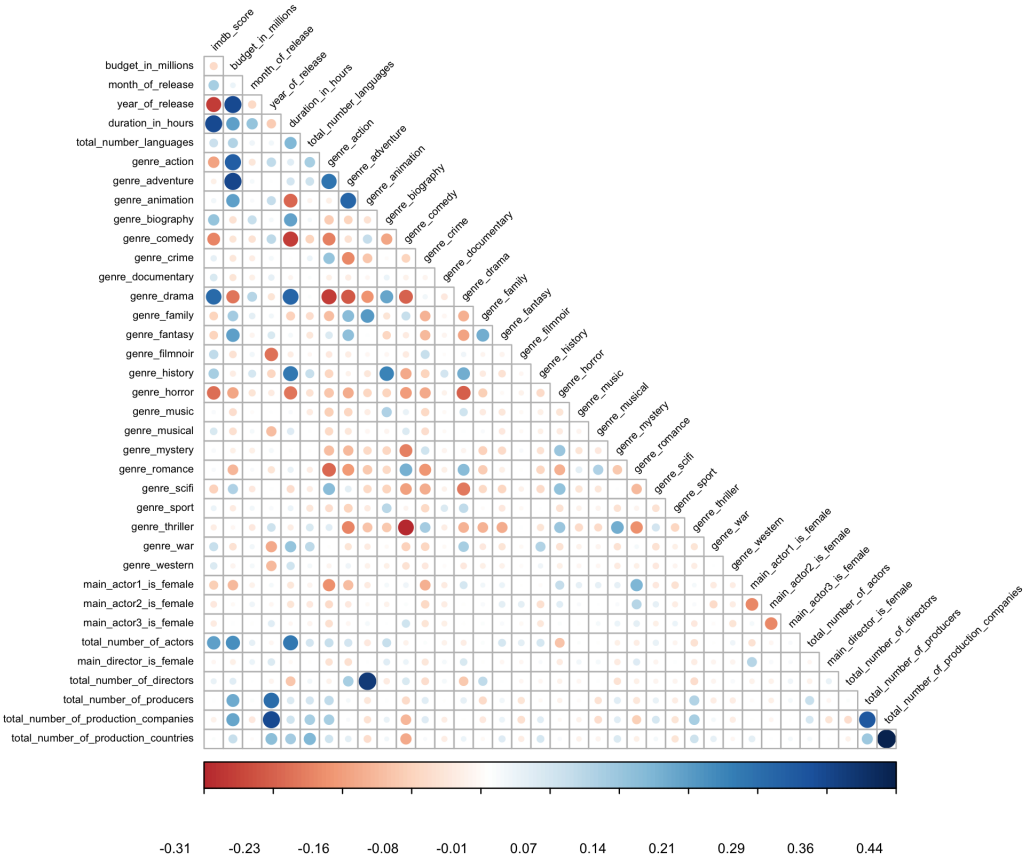

Figure 4: Correlation grid

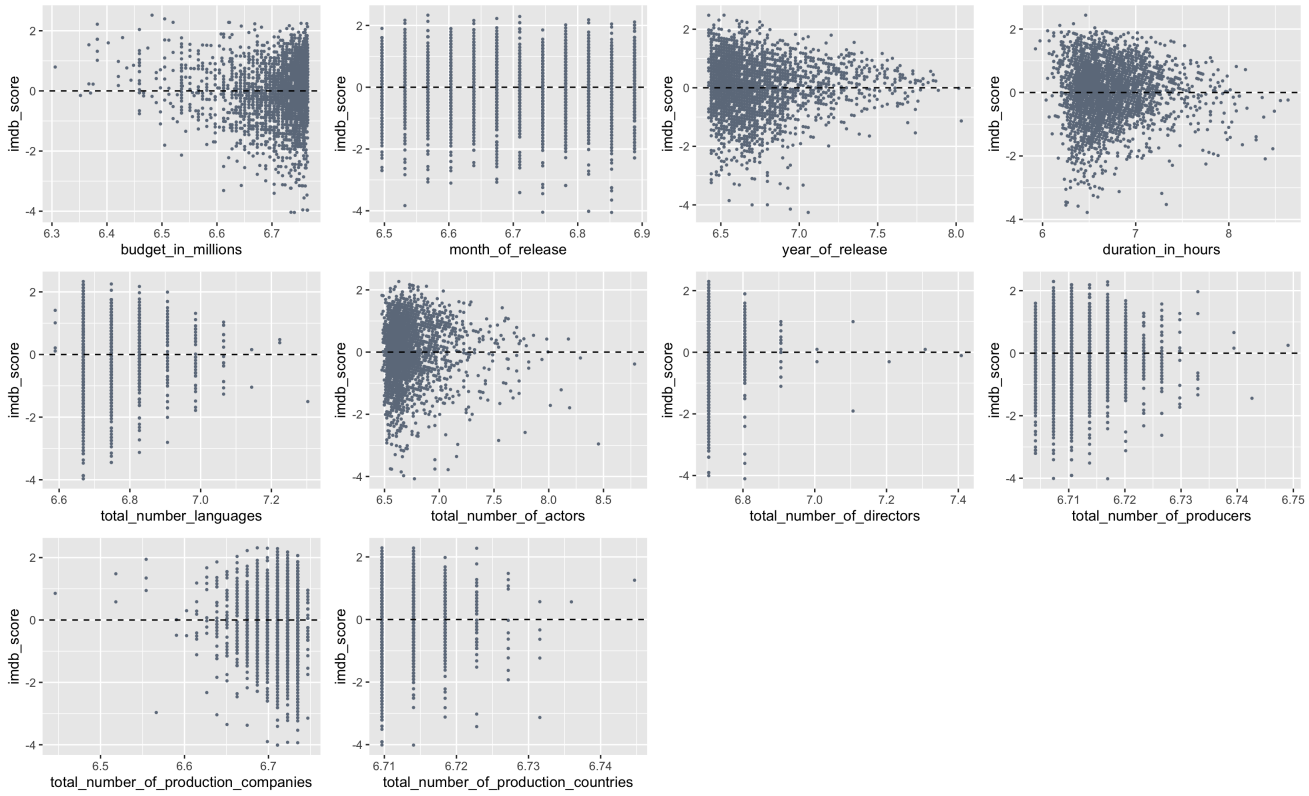

Figure 5: Residual Plots

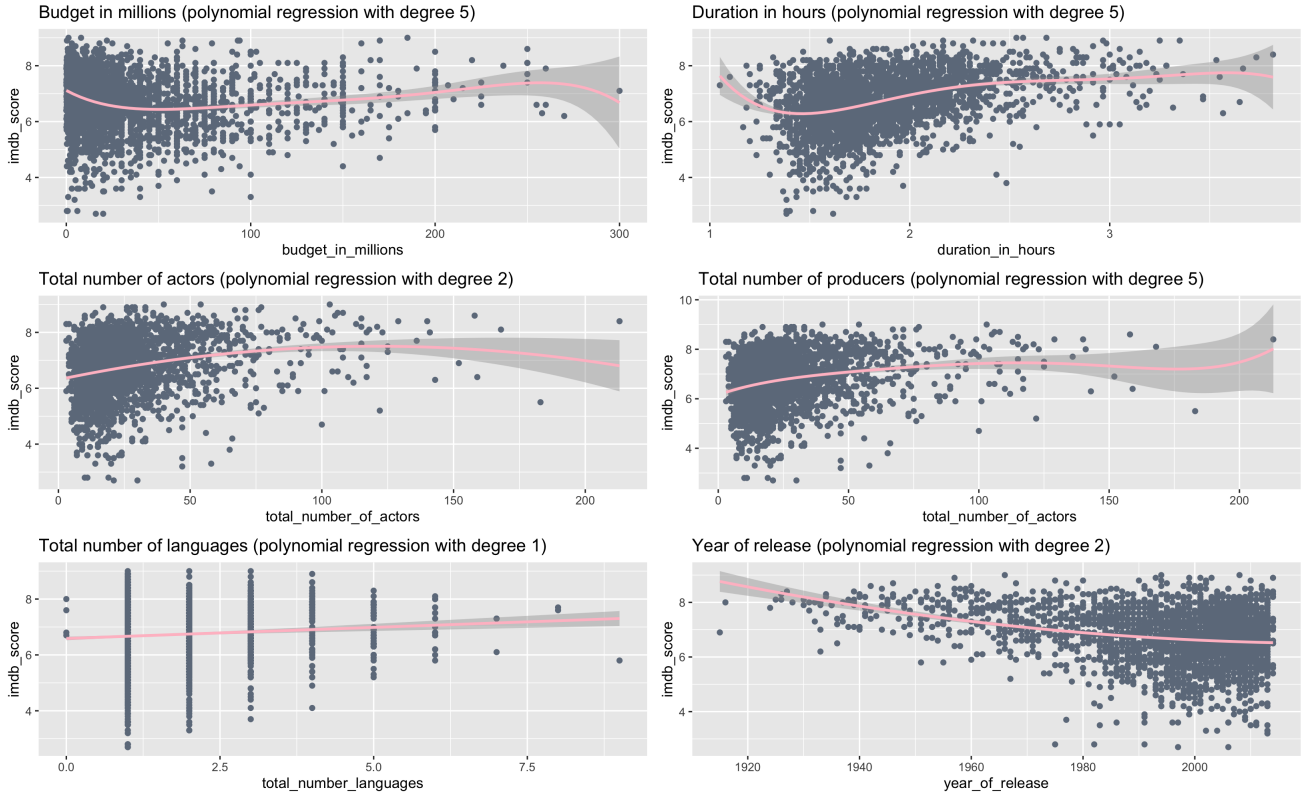

Figure 6: Polynomial regressions grid

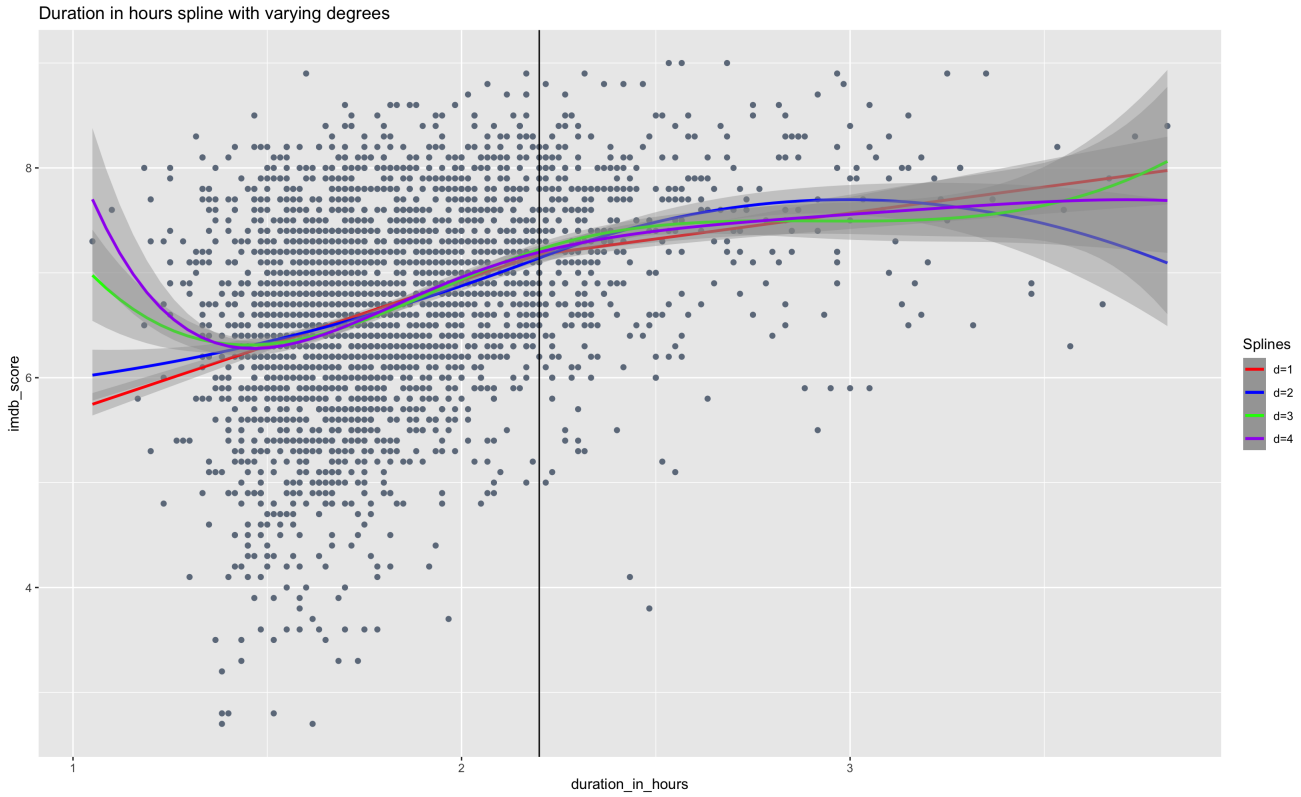

Figure 7: duration_in_hours spline with varying degrees

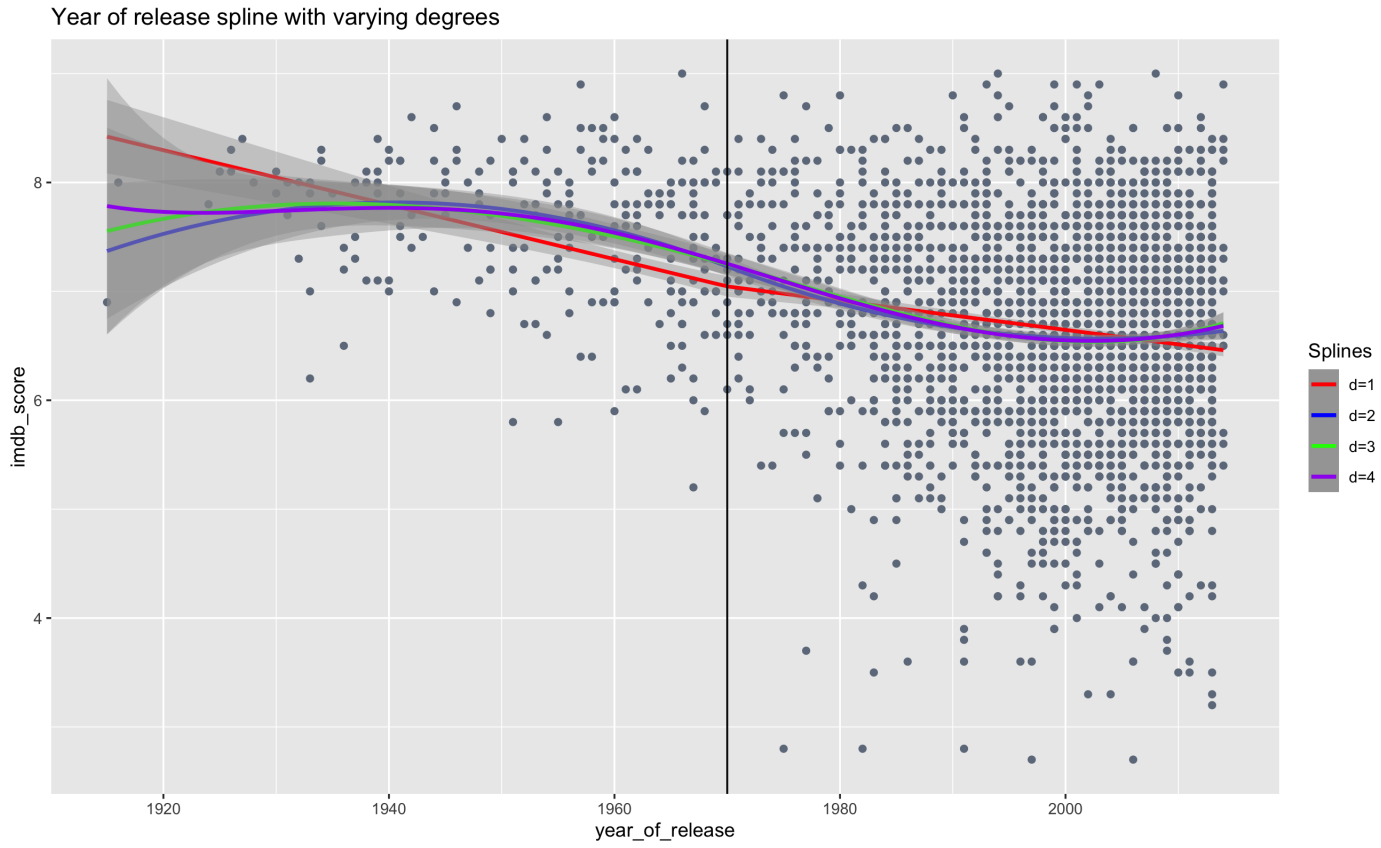

Figure 8: year_of_release spline with varying degrees

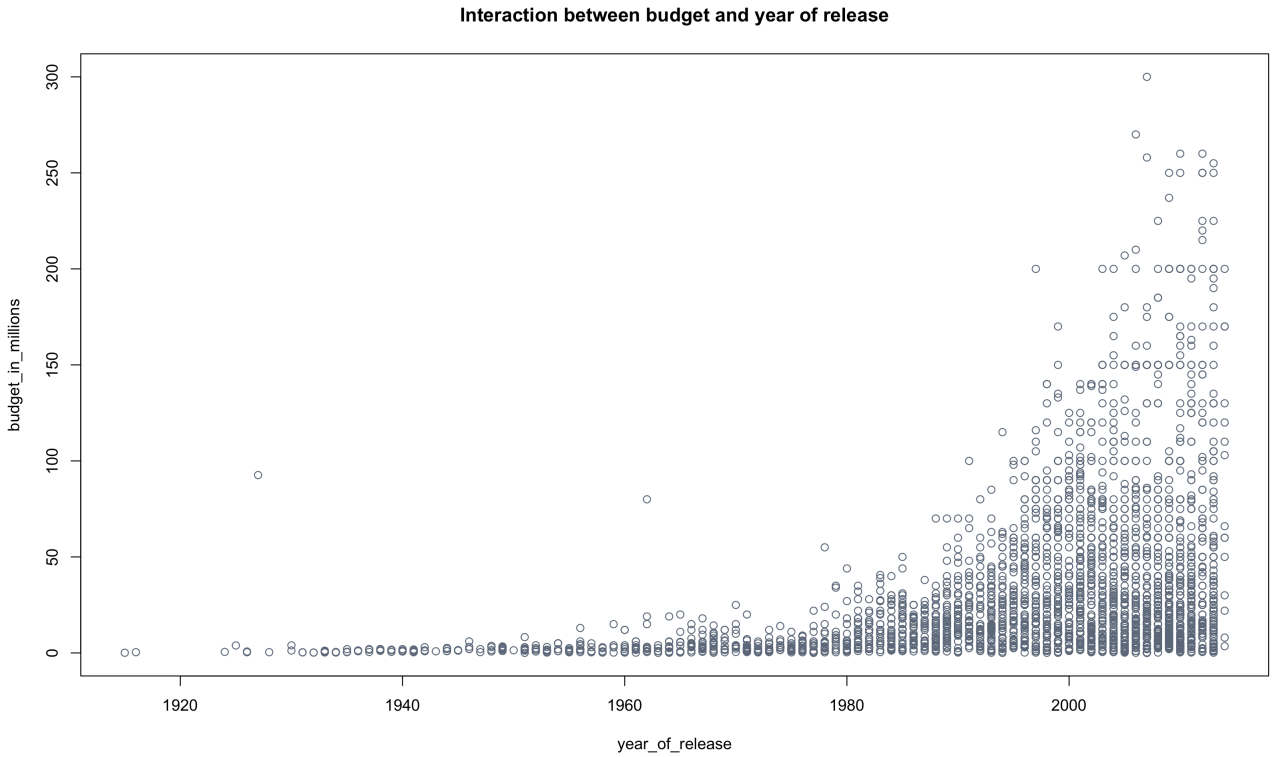

Figure 9: Scatter plot of budget_in_millions and year_of_release

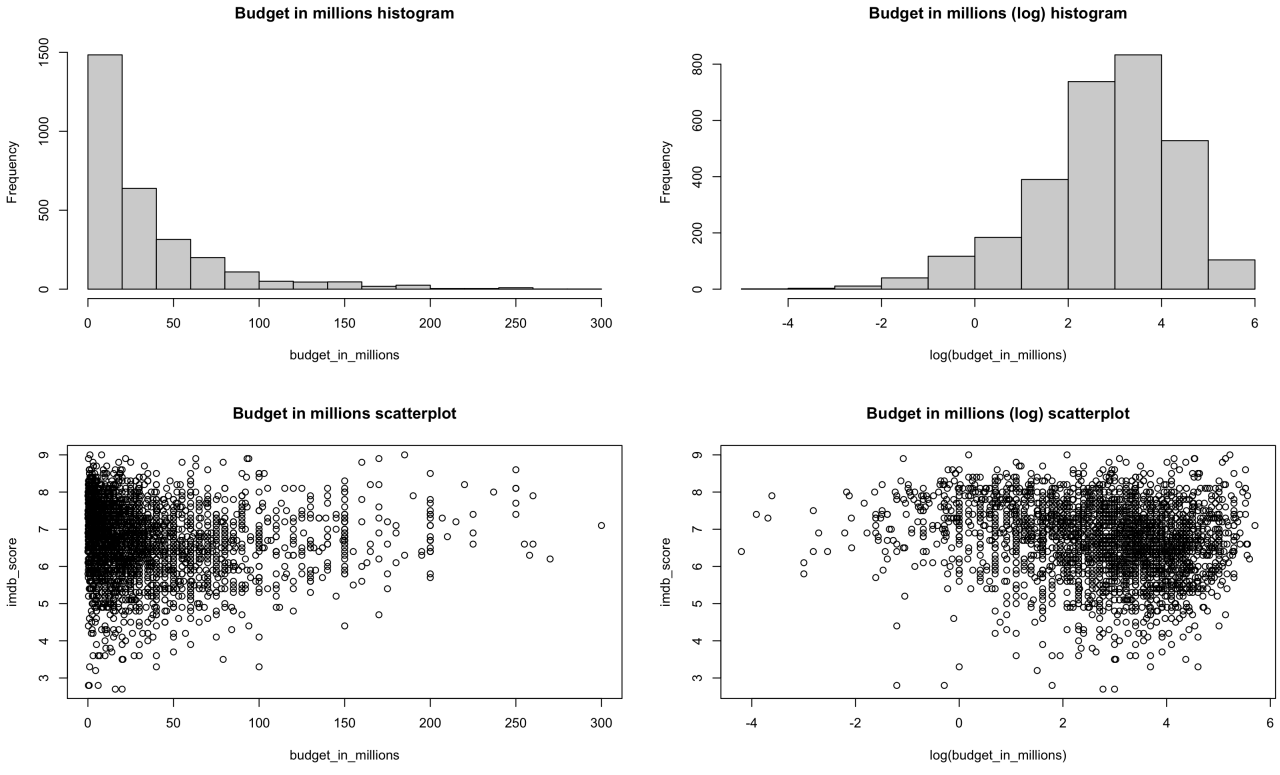

Figure 10: Log transformation of budget_in_millions

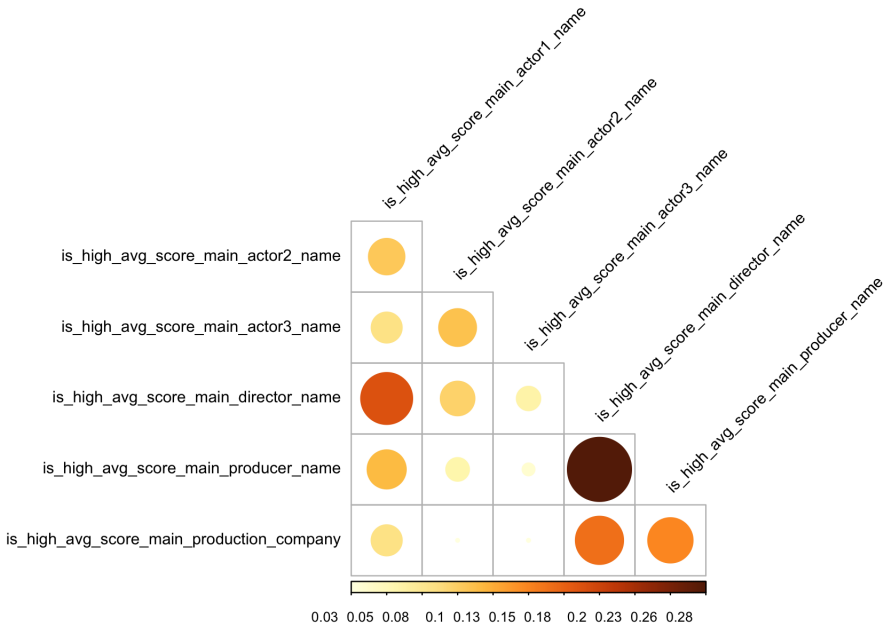

Figure 11: Correlation grid for computed variables