MGSC 401 - Midterm

By: The Standard ~~Error~~

Si-Yuan Jan, Houchen Li, David Paquette, Zhixuan Ren

Introduction

For this project, we were given a large dataset containing information on movies and their respective IMDb (Internet Movie Database) rating. From there, we were given the task of exploring each variable independently, and then exploring the variable relationships, i.e., the relationship between every independent variable and the dependent one. In this situation, the dependent variable was the IMDb movie rating, and the independent variables were all the characteristics of the movie (e.g., the duration of the movie, the year it was released, etc.). These processes involved finding the distribution of the variables, searching for outliers and for the presence of heteroskedasticity, and testing out the linearity assumption for every variable, amongst many other tasks. Ultimately, the main goal for this project was to use all these findings from the exploration of the variables in order to create the best regression model possible, i.e., the one that would give us the lowest mean squared error when predicting the IMDb score of any given movie. By finding the regression model with the lowest mean squared error, we would essentially be finding the line that best fits the dataset, thus enabling us to more accurately predict the IMDb rating of a certain movie. This report will therefore discuss the aforementioned descriptions of variables and their respective relationships with the independent variables, as well as the methodology used to build the model, and finally, the results of our model and the implications of said results.

Data Description

As mentioned in the introduction, the dataset used was a collection of movies listed on IMDb, along with their ratings and the other notable characteristics of the movie. The dataset included 3987 observations. Each observation had 23 different variables. The first two are labels, meaning that they hold no predictive power. These included the title of the movie, and the link to the movie's page on IMDb. After, there was the dependent variable, i.e., the variable we would attempt to predict with our model: the IMDb score of the movie. Following that came all the predictors. The quantitative predictors included: the year the movie was released, the duration of the movie in minutes, the budget of the movie in its local currency (i.e., the currency of the country which the film is native to), the aspect ratio, the number of faces displayed on the movie's poster, the number of likes the movie had on its Facebook page in 2018, the number of

likes the main actor, secondary actor, and tertiary actor had on their Facebook page, and finally, the number of likes the director had on their Facebook page. On the other hand, the qualitative variables included: the main genre of the movie, the secondary genre of the movie, keywords associated with the plot, the language in which the movie was originally released in, the country in which the movie was originally released in, the movie's content rating, the names of the main, secondary, and tertiary actors, and at last, the name of the director of the movie.

Quantitative Predictors

First, the quantitative variables were explored individually in order to understand their distribution. It was important to analyze this as having data that is heavily skewed could greatly affect the coefficients of the regression, or falsely imply that the relationship of a predictor with the dependent variable was statistically significant. After exploring them individually, the relationship between each respective independent predictor and the dependent variable was then examined in order to search for a statistically significant relationship, and to examine the correlation with each other.

The first variable explored was the dependent variable, i.e., the IMDb score of the movie. From the histograms, one can clearly see that the data was left-skewed, suggesting that the dataset contained more movies that scored lower than ones that scored highly. The mean score was 6.456, and the median was 6.600.

Moving on, all the independent quantitative variables were explored. The following list summarizes the findings:

- Year of release: The data is significantly left skewed as most of the movies were released post 1995. The relationship between this variable and the dependent one suggests that it is statistically significant as it possessed a p-value of less than 2e-16. The r-squared suggests little correlation as it holds a value of approx. 0.039.

- Duration of the movie in minutes: The data here is slightly right skewed. The mean movie duration was 109.7 minutes, while the median was 105.0 minutes. The relationship with the score of the movie was statistically

significant. This variable had an r-squared value of 0.1386, suggesting some mild correlation.

- Budget in local currency: The histogram showed the data was extremely right skewed. Furthermore, this variable had a p-value greater than 0.05, suggesting it did not have a statistically significant impact on the rating, as well as insignificant r-squared.

- Aspect ratio: The data had one outlier causing the data to skew further to the right than otherwise expected. When exploring the relationship with the dependent variable, the p-value was 0.269, suggesting no relationship between the two.

- Number of faces on movie posters: The data here is right skewed. Like in the previous variable, there is the case of an outlier heavily skewing the data; however, the relationship did seem to be statistically significant, but the r-squared was extremely little.

- Likes on movie's Facebook page in 2018: The data here was very right skewed; however, the relationship with the dependent variable is statistically significant, and possesses an r-squared of approx. 0.075.

- Likes on the main actor's Facebook page: The data here is heavily skewed to the right due to outliers. However, although it has an insignificant r-squared, the relationship with the dependent variable is statistically significant.

- Likes on the secondary actor's Facebook page: Like the variable before it, outliers are causing this dataset to heavily skew to the right. On the other hand, it does seem to have a statistically significant relationship with the dependent variable.

- Likes on the tertiary actor's Facebook page: Like its "actor like" variables counterparts, the data faces a similar distribution: it is heavily right skewed

due to outliers, yet the relationship with the dependent variable is still significant.

- Likes on director's Facebook page: The data is again heavily skewed to the right; however, the P-value is very low suggesting it does have an impact on the IMDB score of a movie.

Please refer to appendix 1 in order to see the histograms mentioned above, and to appendix 2 to see the scatterplots with regression lines examining the relationships between the dependent and the independent variables.

Qualitative Variables

In terms of qualitative variables, a different approach was employed in order to analyze and explore. Certain predictors (i.e., the names of the main, secondary, and tertiary actors, the directors name, the keywords of the plot) were not analyzed at all due to the fact that they had too many unique values. Consequently, the remaining predictors were dummified and then explored. The following is a summary of the findings:

- Main genre: There are 18 main genres. The most popular genres are comedy, action, drama, adventure, crime, biography, and horror. 99% of the movies have one of the previously listed genres as their main genre.

- Secondary genre: There are 22 secondary genres in the dataset. The most popular ones are drama, adventure, crime, comedy, romance, and mystery.

- Language: While there are 34 languages, 3817 of the 3897 movies available in the dataset were originally released in English.

- Country: The dataset contains movies native to 52 countries. The majority of the films in the dataset, i.e., approx. 80%, were released in the USA.

- Content Rating: There are 15 different content ratings. They range from G all the way to unrated. The most popular content ratings were R, PG-13, and PG. These three made up approximately 92% of all the content ratings.

Model Selection

After plotting data and assessing the distribution of each variable, we first chose multiple numerical variables for model building in order to avoid underfitting issues, and we removed a few **outliers** after visualizing the data (Appendix 3). We also removed certain predictors that seemed rather insignificant or whose R-squared were negligible, such as “aspect ratio” and “number of faces on poster", for we thought that they were not relevant to our prediction model. Moreover, we did not include the predictor “budget local currency” in our model because we thought that the difference in exchange rates would be troublesome, as we would need to find a way to convert them all into the same currency for it to be a fair predictor. For example, a Japanese movie called “Harlock: Space Pirate" has a budget local currency of ¥30,000,000 JPY, that is $2,734,005USD, which makes a huge difference, and it also shows this predictor is meaningless. Furthermore, this predictor would not take into account inflation, suggesting that the purchasing power of older budgets would be severely undervalued, and as a result, unfairly penalized. Excluding these variables helped us avoid overfitting issues since we did not want to add too many useless variables as it would result in a poor out-of-model performance.

Then, we tried to identify model issues including collinearity, non-linearity, heteroskedasticity and outliers. We created a **correlation matrix** to check the correlation between selected variables. The results showed that there was no collinearity between variables, as the correlation coefficients between each pair of variables was less than 0.8. (Appendix 4)

Next, we ran a non-constant variance test and a visual test to detect **heteroskedasticity** in each variable**.** The variables with a p-value of less than 0.05 were noted. Then, we conducted a coeftest in order to eliminate the heteroskedasticity from our model, thus obtaining the real linear significance of each variable.(Appendix 5)

We then built a linear model using selected variables to test the **non-linearity** of each variable. We used residual plots for visualizing the data, and the ResidualPlots function on Rstudio to verify if p-value was less than 0.05. The residual plot revealed a certain pattern, which allows us to conclude that the whole model was not linear. Moreover, we also noted the variables whose p-value was less than 0.05 as these were the ones we needed to explore further for their polynomial degree. (Appendix 6, Appendix 7)

After that, we began finding the polynomial **degrees** for all variables which were determined to be non-linear. For each variable, we tested the polynomial degrees 1 through 5 by using an **anova test** and ggPlot. In this process, we narrowed down the possible degrees for each non-linear predictor and selected several possible models. After running a **K-folds cross validations test** with K=50, we obtained a model with a satisfactory MSE.

With this new model, we rechecked for **new outliers** by using the function outlierTest (Appendix 8); consequently, the new outliers were removed from the dataset.

Next, we began experimenting with dummy variables to add to the model. After plotting the relationship between IMDb scores and the **categorical variables**, we chose five of them to further explore. The categorical variables chosen were: main genre, secondary genre, country, language and content rating. (Appendix 9). By visualizing the data, we chose several dummy variables (e.g., Biography from main genre, UK from country, etc.) from each categorical variable. The final decision was made using K-fold tests and the summary function, by checking the MSE and the adjusted R-squared to see if the dummy variables made significant improvement to the model.

After choosing the dummy variables, we also decided to check if using **splines** could improve the model. As there was no clear visual evidence of where and how to put the knots for the variables, we placed the knots uniformly by the **percentiles** of the data. We tried **3 knots** and **4 knots**. The 3-knots plan worked better, and in the K-fold test, it decreased the MSE significantly. Thus, we decided to include splines in our model. However, for the variable "the number of likes on the main actor's Facebook page", we used a polynomial model because by looking at the relationship between IMDb score and said variable, we thought that a spline would not help much. By performing the test, we confirmed that the results aligned with our thoughts.

Following that, we used a **loop** to recheck the polynomials of each non-linear variable. We chose a low value of K=5 to ensure the code is able to run smoothly. After running the four loops, we found that the degrees of the numerical variables are not exactly the same as the ones we found originally, but they were similar. Therefore, we played around with the numbers in order to find the optimal combination, i.e., the one that produced the smallest MSE. After

multiple trials, we selected the degree of each variable, and we found out that the variable “the number of likes on the secondary actor’s Facebook page " does not significantly decrease the MSE. Therefore, we removed this predictor from our model.

In addition, we tried some interaction terms in our model, but none of them significantly improved the model, so we did not include them in our model. We also retested the heteroskedasticity issue after we obtained the final model and corrected it.

After all these were done, we had our final model:

Numerical Predictor Used	Knots	Degree
Duration in Minutes	3	4
Movie Facebook Likes	3	3
Actor 1 Facebook Likes	N/A	3
Actor 3 Facebook Likes	3	2
Director Facebook Likes	3	4
Title Year	3	3

Table 1: Numerical Predictors Used in the Final Model

Dummy Variable Used	
Language	English
Main Genre	Biography, Action, Horror, Drama,Crime
Content Rating	R, PG-13
Second Genre	Animation
Country	UK

Table 2: Dummy Variable Used in the Final Model

Managerial Implications

Our team focused on creating the most accurate predictive statistical model possible. We noticed that polynomial splines yield a smaller mean-squared error (MSE) than other regression models for most quantitative predictors we used. Hence, we relied heavily on polynomial splines to best predict the IMDb score of a given movie. However, using polynomial splines come at the expense of inference power; meaning that affected predictors can no longer be interpreted. Precisely, 1) *Movie duration*; 2) *Number of likes on Movie's Facebook page in 2018*; 3) *Number of likes on Tertiary actor's Facebook Page*; 4) *Number of likes on Tertiary actor's Facebook Page*; and 5) *Release year of movie* are all quantitative predictors that cannot be interpreted because their coefficients held no meaning in our model. Nonetheless, we can still plot the graphs of each individual predictor to interpret the patterns of the spline polynomials (Appendix 10). By visualizing the matrix of graphs, we can tell a movie director that film rating is generally higher when 1) *Movie duration* is longer than approximately 90 minutes. As for the 2) *Number of likes on Movie's Facebook page in 2018*, 3) *Number of likes on Tertiary actor's Facebook Page* and 4) *Number of likes on Director's Facebook Page*, more likes on these Facebook pages indicate that it should lead to a better film rating. Although an uncontrollable factor, it should be noted that older movies tend to score better, so when 5) *Release year* is smaller.

As for the remaining quantitative predictor, 6) *Number of likes on main Actor's Facebook Page*, the cubic regression graph (Appendix 10) between 'IMDb Score' and the aforementioned variable suggests that movie rating is generally higher when the main actor has around 25,000 thousand likes on Facebook. When an actor has considerably more or less than 25,000 Facebook likes, the IMDb score tends to decrease. Based on this information, we can assume that the director should hire a sufficiently popular main actor with around 25,000 Facebook likes in order to potentially achieve maximum film rating.

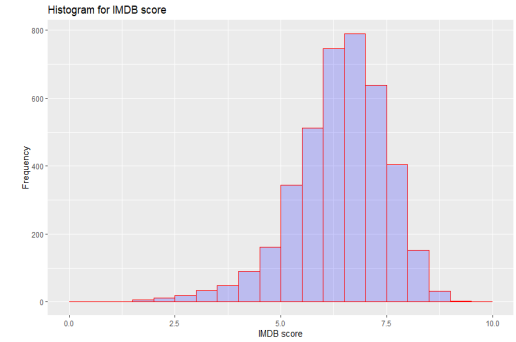
Finally, we have to consider the importance of dummy variables. Before interpreting any data, it is crucial to note that all the coefficients to be presented thereafter consider the relationship between each predictor in the model. In other words, the coefficient values are interdependent between all variables. A film director has to utilize our model in its entirety to justify the prediction he generates, and not just use each dummy variable separately or else he

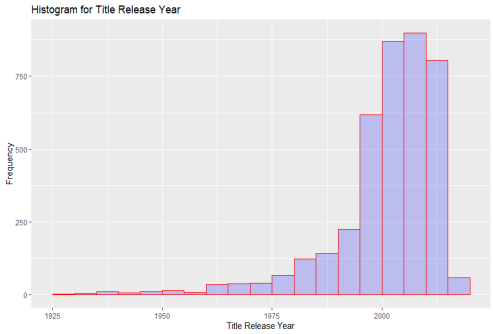
will get inaccurate predictions. A dummy variable can only take the value of either 0 or 1 in our model, which indicates the absence or the presence of a specific condition that will either unaffect or affect our IMDb score prediction. Simply put, does a movie's qualitative characteristics influence the rating of a film? Our regression model includes five qualitative predictors, which are represented by ten dummy variables that we believe have an impact on the movie score. Inferences are made by interpreting the coefficients of our model (Appendix 11).

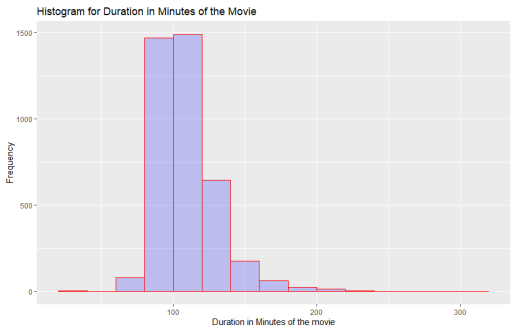
First, if the movie's language is in English, then our model suggests that the rating will generally decrease by around 0.589. This is because the majority of movies in the sample size are in English, so there are more instances of bad movie scores compared to other languages. Thus, our model will automatically give a lower score to movies in English. Second, we noticed that a few primary genres significantly affect film ratings. Specifically, *biography* (+0.469); *action* (–0.237); *horror* (–0.584); *drama* (+0.253); and *crime* (+0.248) genres tend to impact the film rating. Our guess is that some main genres achieve a better score because they are unsaturated (e.g. biography) so there are less bad movies, they are easier to produce, or they are naturally more enjoyable. Likewise, when a film's secondary genre is labelled as an animation, the IMDb score seems to rise by 0.555, so a director should think about producing an animation movie regardless of the main genre. Third, the content rating also seems to influence the movie score. Indeed, movies that are labeled as R tend to have a better movie score of 0.174, and movies labeled as PG-13 lower the score by 0.129. R and PG-13 movies are the most common in our data set, and R movies are generally rated higher than PG-13 despite having more observations. Hence, this could explain the coefficients of our model. A film director should avoid PG-13 movies and produce a R-rated movie instead if they wish to maximize the IMDb score. This occurrence can be explained by the fact that R movies do not compromise on content censorship, so the films are more authentic and genuine. Finally, a movie generally has a better rating if it is produced in the UK. This country speaks English, which is the most common language found in this dataset, yet UK movies tend to score higher than many other countries (including the US). Precisely, British films typically score higher by 0.189 compared to other nations. This is probably because the UK produces fewer movies than Hollywood, so UK directors tend to make every movie count. Contrarily, the US produces more movies (quantity), so there are a number of poorly rated movies that skew the data unfavorably.

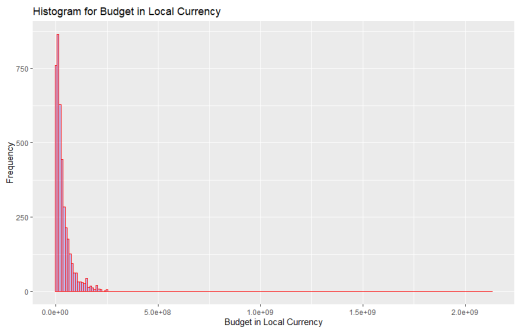
APPENDICES

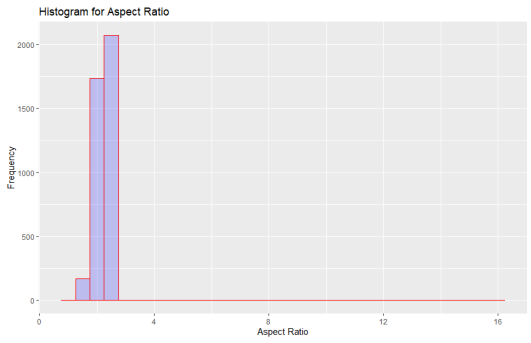
Appendix 1: Distributions of Predictors

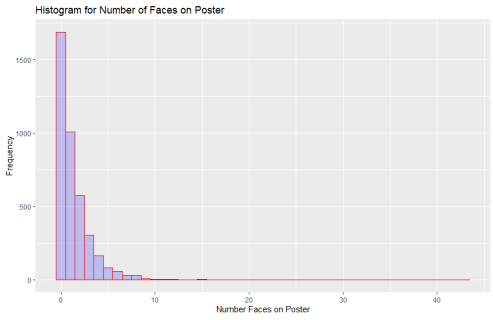


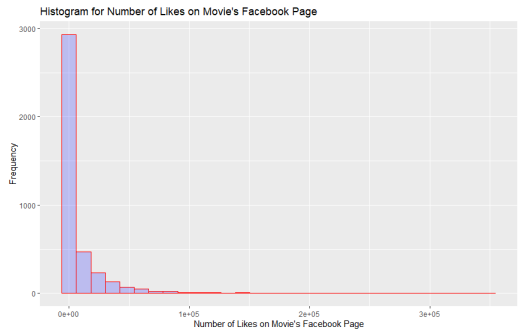


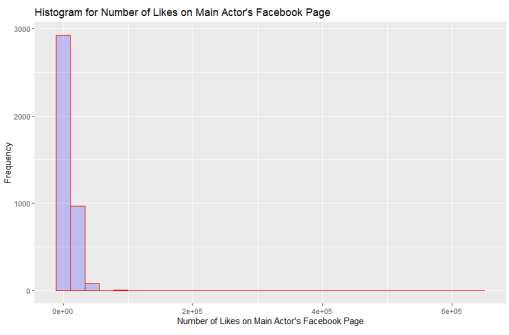


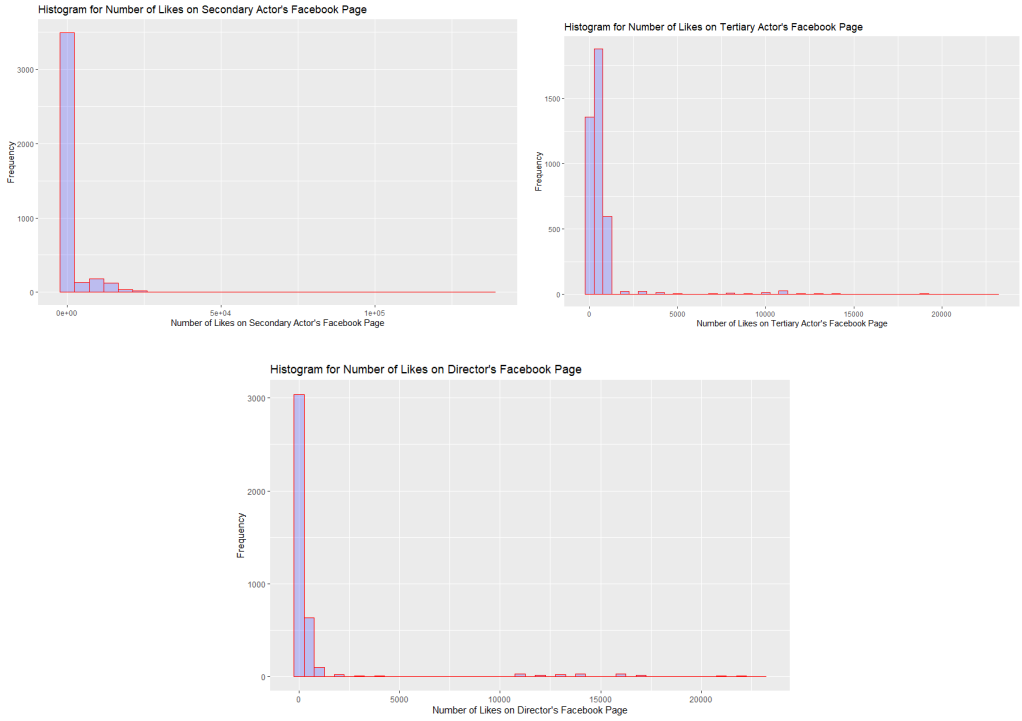




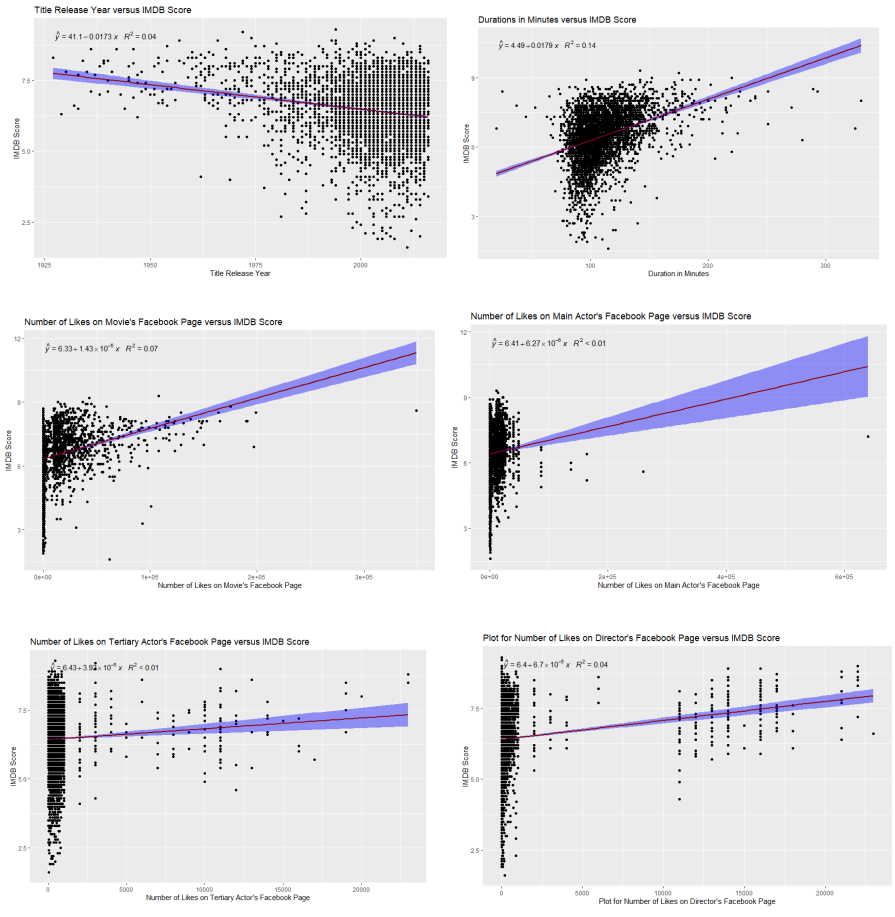




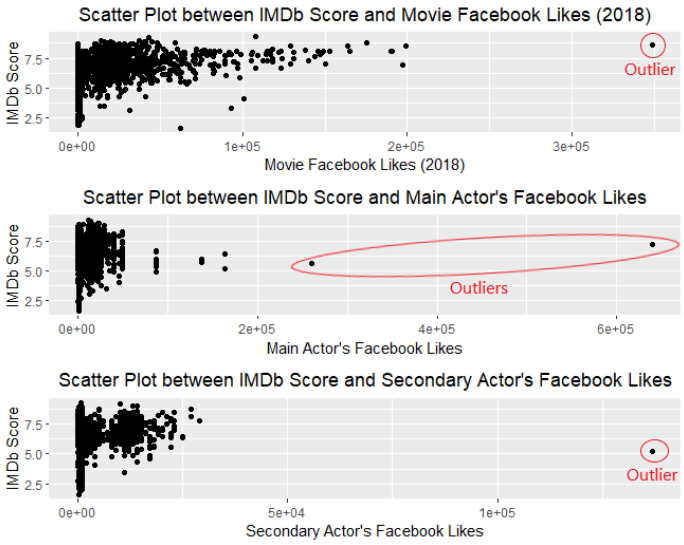




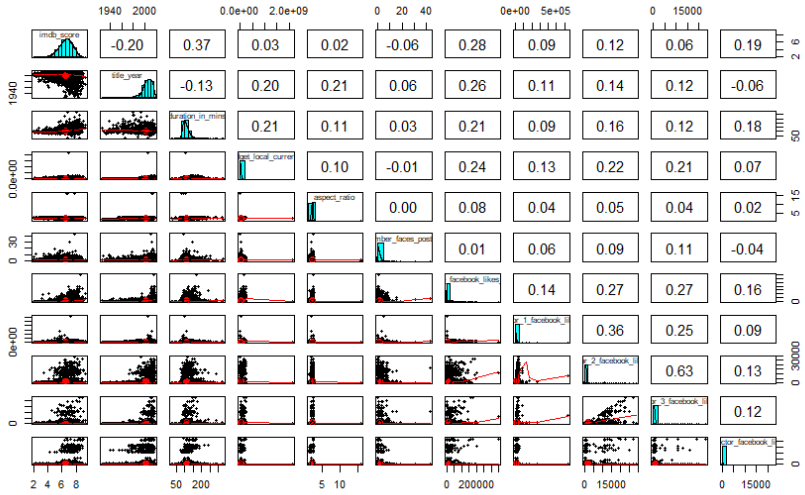
Appendix 2: Plots Plots of IMDb Scores vs. Quantitative Variables



Appendix 3: Visual Outlier Test for Individual Quantitative Predictors



Appendix 4: Correlation Matrix of Predictors to Test Collinearity



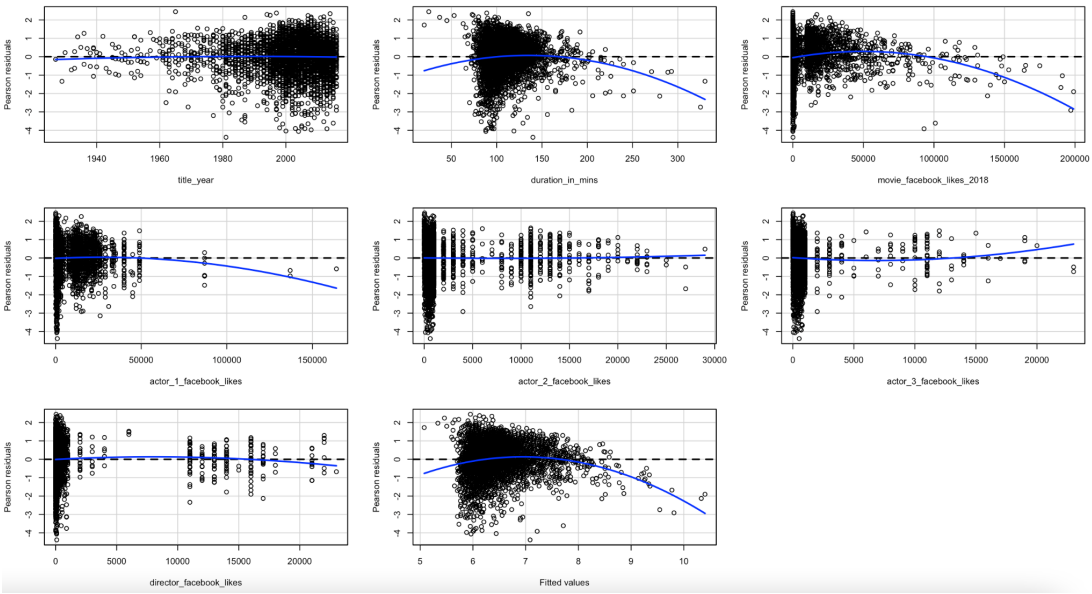
Appendix 5: Non-Constant Variance Test P-values to Detect Heteroskedasticity

Non-constant Variance Test	
Variable Name	**NCV p-value**
Movie Facebook Likes	0.43649
Actor 1 Facebook Likes	0.63913
Actor 2 Facebook Likes	0.36723
Actor 3 Facebook Likes	0.011882
Director Facebook Likes	2.3041e-06
Title Year	0.027213
Duration in minutes	2.9822e-12

Appendix 6: Residual Plot P-values of Predictors to Test Linearity

```
> residualPlots(mreg)
                          Test stat Pr(>|Test stat|)
title_year                  -0.9379         0.348377
duration_in_mins            -6.4214         1.51e-10 ***
movie_facebook_likes_2018  -10.5425        < 2.2e-16 ***
actor_1_facebook_likes      -3.0844         0.002053 **
actor_2_facebook_likes       0.5687         0.569580
actor_3_facebook_likes       2.1919         0.028446 *
director_facebook_likes     -1.7318         0.083392 .
Tukey test                 -10.6122        < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

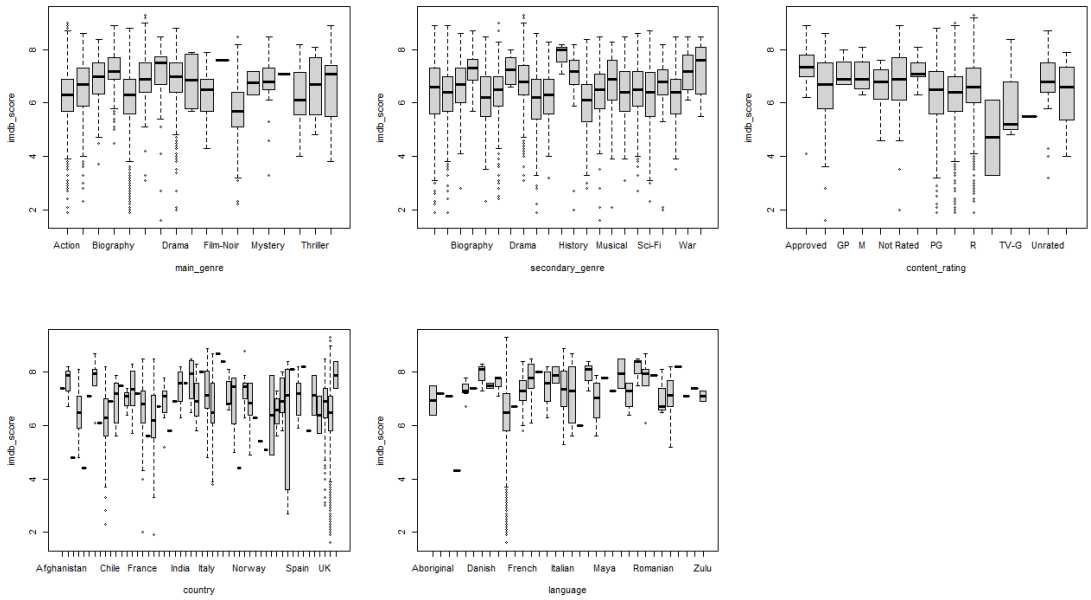
Appendix 7: Residual Plots of Predictors to Test Linearity



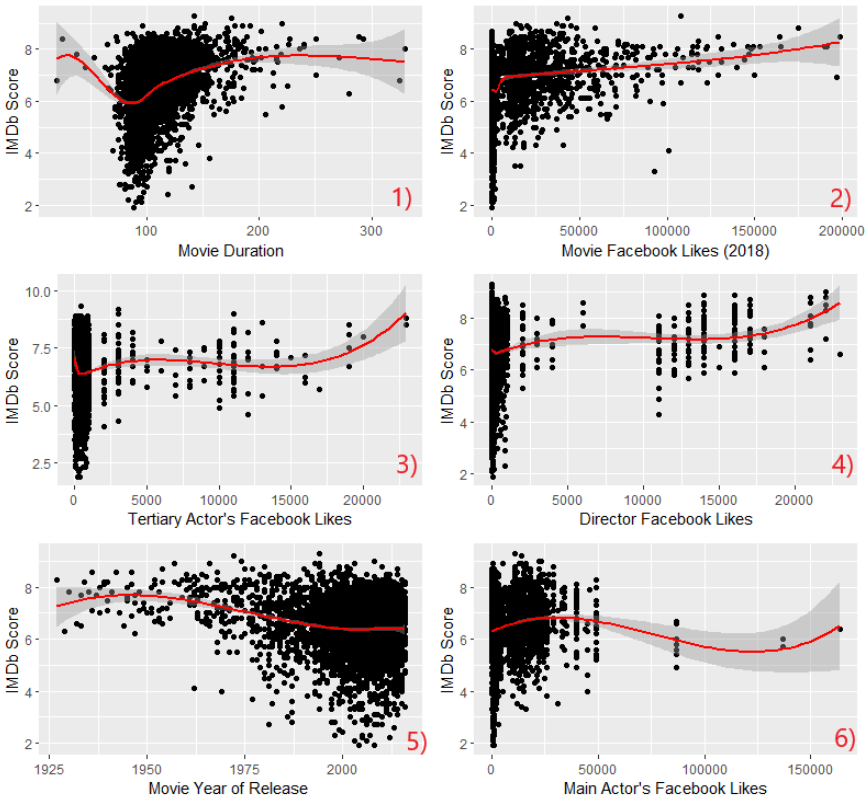
Appendix 8: Bonferroni Outlier Test for our Final Regression Fit

```
> outlierTest(mreg2)
      rstudent unadjusted p-value Bonferroni p
1594 -7.295990         3.5865e-13   1.4135e-09
3435 -5.094181         3.6717e-07   1.4470e-03
1184 -5.045345         4.7375e-07   1.8671e-03
3776 -5.031340         5.0946e-07   2.0078e-03
1191 -4.853165         1.2636e-06   4.9797e-03
3134 -4.716482         2.4852e-06   9.7943e-03
3901 -4.443633         9.0958e-06   3.5846e-02
1681 -4.381499         1.2103e-05   4.7697e-02
```

Appendix 9: IMDb Scores vs. Categorical Variables to Select Dummy Predictors



Appendix 10: Regression Plots of our Final Model's Quantitative Variables



Appendix 11: Coefficients of Final Model

```
Coefficients: (1 not defined because of singularities)
                                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                                      11.05917    0.74940  14.757  < 2e-16 ***
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)1             0.56224    0.85850   0.655 0.512566
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)2            -1.52912    0.55209  -2.770 0.005637 **
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)3            -0.55002    0.56912  -0.966 0.333889
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)4            -0.18873    0.57686  -0.327 0.743564
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)5             0.04791    0.73684   0.065 0.948160
bs(duration_in_mins, knots = c(ki_1, ki_2, ki_3), 4)6            -0.35074    0.74635  -0.470 0.638429
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)1 -1.84904    0.39233  -4.713 2.53e-06 ***
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)2 -2.72108    0.39261  -6.931 4.87e-12 ***
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)3 -1.40147    0.40575  -3.454 0.000558 ***
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)4 -0.07203    0.37094  -0.194 0.846044
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)5 -0.41929    0.70980  -0.591 0.554748
bs(movie_facebook_likes_2018, knots = c(kj_1, kj_2, kj_3), 3)6       NA         NA      NA       NA
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)1    -0.44369    0.11157  -3.977 7.12e-05 ***
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)2    -0.52945    0.07648  -6.922 5.16e-12 ***
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)3    -0.53661    0.07195  -7.458 1.08e-13 ***
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)4    -0.49268    0.35980  -1.369 0.170972
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)5    -1.38015    0.54351  -2.539 0.011144 *
bs(actor_3_facebook_likes, knots = c(km_1, km_2, km_3), 2)6     0.35236    0.50208   0.702 0.482848
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)1   -0.18252    0.06457  -2.827 0.004729 **
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)2   -0.22907    0.05129  -4.466 8.19e-06 ***
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)3   -0.03535    0.03933  -0.899 0.368787
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)4    0.68634    0.38629   1.777 0.075685 .
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)5   -0.18794    0.39396  -0.477 0.633350
bs(director_facebook_likes, knots = c(ko_1, ko_2, ko_3), 4)6    0.36279    0.27095   1.339 0.180659
poly(actor_1_facebook_likes, 3)1                                2.55011    0.87356   2.919 0.003529 **
poly(actor_1_facebook_likes, 3)2                               -1.14201    0.80242  -1.423 0.154754
poly(actor_1_facebook_likes, 3)3                                1.76085    0.77954   2.259 0.023949 *
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)1                 0.75380    0.52961   1.423 0.154728
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)2                -1.07709    0.25730  -4.186 2.90e-05 ***
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)3                -0.90280    0.29664  -3.043 0.002354 **
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)4                -1.24530    0.28281  -4.403 1.09e-05 ***
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)5                -2.18354    0.30250  -7.218 6.29e-13 ***
bs(title_year, knots = c(kq_1, kq_2, kq_3), 3)6                -1.54807    0.29410  -5.264 1.49e-07 ***
lan_eng                                                        -0.58929    0.06581  -8.954  < 2e-16 ***
main_g_bio                                                      0.46856    0.05734   8.172 4.04e-16 ***
main_g_action                                                  -0.23664    0.03216  -7.358 2.27e-13 ***
main_g_hor                                                     -0.58417    0.06126  -9.536  < 2e-16 ***
main_g_dra                                                      0.25253    0.03605   7.005 2.89e-12 ***
main_g_cri                                                      0.24787    0.05250   4.721 2.42e-06 ***
content_R                                                       0.17408    0.03617   4.813 1.54e-06 ***
content_PG13                                                   -0.12977    0.03834  -3.385 0.000719 ***
sec_animation                                                   0.55543    0.07677   7.235 5.56e-13 ***
country_UK                                                      0.18970    0.04339   4.372 1.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.5708117)

    Null deviance: 4540.1  on 3970  degrees of freedom
Residual deviance: 2242.1  on 3928  degrees of freedom
AIC: 9087.5

Number of Fisher Scoring iterations: 2
```