MDSA 603 - Final group project (Group 1)

Contents

1 Introduction

The global entertainment industry has an estimated market capitalization(Market Reports World 2022) of \$3.5 trillion USD [1]. \$90 billion of which are on movies alone(Grand View Research 2022). Seeing as this is a larger number than more than one country's GDP, we're guaranteed some interest in statistical models of a

[1] Using the american definition of billion/trillion

popular industry. Other than this fact there is little relevance to society as a whole. Another toy model that helps us-students- progress in our understanding of multi-linear statistical modelling.

Our group members,

- Ali Raza (UCID 30084838)
- Orlando Morales (UCID 30190412)
- Roberto Sierra Ortega (UCID 30168140)
- Stuart Finley (UCID 30191070)
- Jose Palacios (UCID 30190988)

2 Methodology

2.1 Dataset

The dataset we chose is a publicly available (for personal and non-commercial use (IMDb 2022)). It covers movies and screen-entertainment released since before 1900. These tables are organized in a way typical of relational databases and linked using a pair of unique keys. See Fig. (1) for a high-level view of the dataset's structure.

Originally we wanted to look at revenue per movie. As in many other large for-profit endeavors, budgetary and cash-flow information is highly guarded or simply not accessible to the general public or academics. We decided to drop this idea and wrangle other predictors into existence instead.

By means of data wrangling we created new predictors such as,

- Title word count
- Number of releases in other formats/languages
- Total number of movies released on the same year any given movie was released

These and other predictors are discussed in the following section.

2.2 Licensing

The dataset(IMDb 2022) is published by its owner and it's made available from non-commercial/personal use. They do offer a tier for commercial applications and possibly ready-made models using this data.

2.3 Data Cleaning and Manipulation

The master dataset was loaded into a jupyter notebook along with the subtables. Initial minor cleaning of these datasets was required to convert the NA type used in the dataset to a standard NA recognized by R. The sub tables all include a t-const variable, which is the key that can be used to link all the tables together. The tables were merged together using this key on left merges to generate the single dataframe to be used in the modeling process.

Originally, the dataset consisted primarily of categorical data such as movie type, genres and if the movie is an adult film or not. Only three numerical predictors existed - average ratings, runtime and number of total votes for rating of a film. Anticipating we would require numerical predictors in addition to the categorical predictors already present to build a suitable model, four numerical predictors were derived from the existing tables. These four numerical predictors were the sum total of movies produced in the same calendar year (releaseCount), the average number of votes for ratings (averageRatingVotes), the number of different versions the movie was released in, the different release languages (titleNumReleases) , as well as the word and char counts of the movie titles.

Additional manipulations were conducted on the genres column. Originally, the genre for a given movie consisted of one or up to any 3 combinations of the aforementioned genres. This column was broken up into three separate columns so that the primary genre of the up to 3 genres could be extracted. The final result was a single column for genres which only contained a single genre. This table was then read into R using the F function “read.csv”.

2.4 Dependant and Independent Variables

The dependent variable in our model is the average rating of movies. The variables that are used in our modeling process for the calculating the average rating of movies are:

- **releaseYear**: The year the movie was released
- **titleNumReleases**: The number of different versions the movie was released in, or different release languages
- **titlewordcount**: The number of words in the movie title
- **titlecharcount**: The number of characters in the movie title
- **runtimeminutes**: The duration of the movie in minutes
- **averageRatingvotes**: The number of ratings the movie received
- **releasecount**: The total number of movie releases in the same calendar year as the movie being modeled
- **genre**: Categorical variable, possible categories of the movie are
  - Action
  - Animation
  - Comedy
  - Crime
  - Documentary
  - Drama
  - Horror
  - other
  - Romance

2.5 Modeling process

The modeling process begins once all the variables have been selected for the model. It is important to verify the absence of multicollinearity, which is where two or more of the independent variables in the model are correlated. If present, multicollinearity among independent variables will result in less reliable results, so independent variables will be removed from the model if it’s deemed there exists correlations. Multicollinearity can be tested using scatter plots between independent variables and the The variance inflation factor (VIF) test. The next step in the modeling process is to use stepwise linear regression to determine which independent variables are most important for prediction. This will assist in removing any variables that don’t carry any statistical significance for prediction purposes.

With the most significant predictors selected, an F-test will be performed as well as individual t-tests to test the null hypotheses that none of the predictors carry statistical significance, and then to test which predictors are statistically significant in the model. Individual t-tests are also used to check the significance of the interaction terms between the independent variables as well as the significance of higher order terms. Significant interaction terms and higher order terms will be kept, and a final model will be established. The threshold used for the F-test and the t-test will be testing the significance against $\alpha = 0.05$.

Once the final model is selected, it must then be tested against 5 conditions. These conditions are as follows:

1. **Linearity** The linear regression model assumes that there is a straight-line relationship between the predictors and the response. Condition can be verified using residual plots and checking for any discernible patterns.

2. **Normality** The error between observed and predicted values should be normally distributed. Can be verified using graphs and the Shapiro-Wilk normality test.

3. **Independence** The error terms are uncorrelated. Condition can be verified using residual plots and checking for any discernible patterns or clustering.

4. **Homoscedasticity** The error terms have a constant variance. Can be verified using the Breusch-Pagan test

5. **Leverage Points/Outliers** Points that fall horizontally far from the line can strongly influence the slope of the least squares line. Influential points can be detected using Cook's distance.

If the final model fails to satisfy any of these conditions, further work is required to determine if any improvements can be made such that the condition is satisfied.

If the linearity assumption is not met, one approach is to use non-linear transformations of the predictors. For independence, since the movie dataset is not a time series dataset, we can infer the measurements are independent. If heteroscedasticity is detected, one possible solution is to transform the response Y using a concave function such as log(Y) or root(X). If leverage points are detected, these observations can be removed and the model can be re-run. If normality is not detected, then the predictor variables can be transformed. However, this is only appropriate if non-linearity is the only condition of the aforementioned conditions which is not met. If a transformation is made, all conditions should be checked again.

2.6 Work distribution

The workload was distributed equitably among the members of our group. All group members took part in the process of exploring the original dataset and determining which variable was to be used as the dependant variable, as well as which variables could be used as independent predictors. Jose primarily did the data cleaning in Jupyter as well as merging the tables together and generation of new numerical predictors based on his proficiency with Python. Jose also wrote the introduction of the report and described the dataset. Orlando and Roberto focused on the results section of the report and built the model in R in addition to testing the required assumptions to validate the model. The other members, Ali, Jose, and Stuart made contributions to the verification of the model while testing the assumptions. Ali and Stuart Wrote the methodology section of the report and contributed to the results and discussion section of the report. All group members assisted in final reviews of the report and in making final changes and adjustments.

3 Results

We built a first order model that consider every variable explained before as a base, and it's as follows:

First order Model:

$$\begin{aligned}\widehat{Y_{Rating}} &= \beta_0 + \beta_1 X_{Release.Year} + \beta_2 X_{Num.Releases} \beta_3 X_{Word.Count} + \beta_4 X_{Char.Count} \\ &+ \beta_5 X_{runtime.Minutes} + \beta_6 X_{Avg.Rating.Votes} + +\beta_7 X_{Release.Count} + \beta_8 X_{genre}\end{aligned}$$

Next, we performed a Full Model Test with significance of $\alpha = 0.05$ to the model obtained from the best subset in order to confirm whether or not we can reject the Null Hypothesis. Here we obtained an $F - statistic = 4997.6$ and a $p - value < 2.2e^{-16}$, which falls well below the significance level.

Based on the high F-statistic and low p-value, we can therefore conclude that we can confidently reject the Null Hypothesis and state that at least one of the predictors is significant to the Rating.

3.1 Variable Selection Procedures

To start with the variable selection and conclude which variable we should include for our model we tried different strategies:

First, we wanted to run stepwise regression but this was not possible because Rstudio kept giving us the error "Error in if (pvals[minp] <= pent) { : argument is of length zero" which we will consider as future work as the workarounds we found online were not able to produce any result; that's when we decided to move on and tried to drop variables based on their VIF values to check if there was any multicolinearilty between our variables and based on those results only one variable got VIF detected but it was one of the values of the categorical variable, hence dropping the categorical variable was not an option.

So, after we didn't succeed to run either stepwise regression or VIF tests, we decided to use best subset in order to try to check if we would need to drop some variables and based on the result of 8 submodels, the one that adapts better for us, meaning that the R^2_{adj} was kind of the same as the others, has a good cp value (model + 1) and the AIC was also similar to the other models, we finally concluded that for our model we should keep 3 variables: Title Char Count, Rating and Genre.

3.2 Hypothesis Statement for Individual t-test:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$
$$\text{(i=title.char.count,Avg.Rating.Votes, genre)}$$

3.3 Main Effects Individual t-test

$$\text{Title Char Count} : t = 15.373, p - value < 2e^{-16}$$
$$\text{Average Rating Votes} : t = 47.438, p - value < 2e^{-16}$$

In order to confirm the right selection of the variables, we used individual t-test with a significance of $\alpha = 0.05$ and as we can see from the results above, we would reject the Null Hypothesis and accept the Alternative, stating that for this model both predictors, Title Char Count and Average Rating Votes, are significant and can be added to the model, concluding that at this point, the model is:

Reduced Model:

$$\widehat{Rating} = \beta_0 + \beta_1 X_{title.char.count} + \beta_2 X_{Avg.Rating.Votes} + \beta_3 X_{genre}$$

3.4 Higher Order Terms

In order to see if the model improve by higher order terms we tried to squared the variables, and the results are as follow:

3.5 Hypothesis Statement for Individual t-test (Higher Order Terms)

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$
$$(i = title.char.count^2, Avg.Rating.Votes^2)$$

We tried to check for high order terms in the model, when we ran the t-test in order to see if they were significant we found that both predictors are significant but when we compared the R^2_{adj} the improvement was of less than 1% and due to complexity and basically not enhance in the model, we decided not to keep those terms

Higher Order Model:

$$\begin{aligned}\widehat{Rating} = \beta_0 + &\beta_1 X_{title.char.count} + \\ &\beta_2 X_{Avg.Rating.Votes} + \beta_3 X_{genre} + \\ &\beta_4 X^2_{title.char.count} + \beta_5 X^2_{Avg.Rating.Votes}\end{aligned}$$

3.6 Higher Order Individual t-test

$$\text{Title Char Count}^2 : t = 8.149, p - value < 3.7e^{-16}$$
$$\text{Average Rating Votes}^2 : t = -21.509, p - value < 2e^{-16}$$

3.7 Interaction terms

Searching for interaction terms to enhance our model, we found that interaction terms are significant for some of the variables, specifically the case for Title Char Count and Average Rating Votes and some interactions between Genre and Title Char Count. That's why we decided to include those interactions in the model, and run individual t-test to confirm if they are still significant.

3.8 Hypothesis Statement for Individual t-test (Interaction Terms)

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

where,

(i=title.Char.count*average.rating.votes, title.char.count*genre, average.rating.votes*genre)

Higher Order Model:

$$\begin{aligned}\widehat{Rating} = \beta_0 + &\beta_1 X_{title.char.count} + \beta_2 X_{Avg.Rating.Votes} + \\ &\beta_3 X_{genre} + \beta_4 X_{title.char.count} X_{Avg.Rating.Votes} + \\ &\beta_5 X_{Avg.Rating.Votes} * X_{genre} + \beta_5 X_{title.char.count} * X_{genre}\end{aligned}$$

3.9 Interaction terms t-test

$$\text{Title Char Count * Average Rating Votes} : t = -5.160, p - value < 2.47e^{-07}$$
$$\text{Title Char Count * genre(Action)} : t = 5.589, p - value < 2.29e^{-08}$$
$$\text{Title Char Count * genre(Animation)} : t = 4.718, p - value < 2.38e^{-06}$$
$$\text{Title Char Count * genre(Documentary)} : t = 4.165, p - value < 3.11e^{-05}$$
$$\text{Title Char Count * genre(Other)} : t = -3.482, p - value < 0.000497$$
$$\text{Title Char Count * genre(Crime)} : t = 3.127, p - value < 0.001766$$
$$\text{Title Char Count * genre(Comedy)} : t = -2.056, p - value < 0.0398$$

As we can observed, the interaction terms seems to be valid for this case as the t and p values are acceptable, but during the t-test we realized that, similar to the case of higher order terms, the interaction terms don't enhance significantly our model, in fact is again less than 1% and for that reason we concluded that is better to keep the reduced model, as something else will only add complexity and not improve the model, that's why, we keep the following model:

$$\widehat{Rating} = \beta_0 + \beta_1 X_{title.char.count} + \beta_2 X_{Avg.Rating.Votes} + \beta_3 X_{genre}$$

4 Multiple Regression Assumptions

For this section, we will test our model to check if it fulfill the requirements based on assumptions that are associated with multiple linear regression. The objective of these tests is to ensure that the model built before is a valid model, or in other words is a model that can make a good prediction based on the R^2_{adj}

4.1 Independence Assumption

When checking for independence of variance of residuals by plotting these by year of release, we observe that they conform into a cone shape. While this would usually indicate that the variance is NOT independent, we believe that is simply influenced by the density of our data in more recent years. This is demonstrated by the following histogram showing the number of films per year in our data set, which shows a clear increase in the number of films particularly since the mid 2010´s.

See Figs. (6) and (7).

4.2 Linearity Assumption

Since one of our base assumptions our model is that there is a linear relationship between our dependent and independent variables. In order to confirm this, we plotted the residuals of our best model, so that we might discern the existence of any non-linear patterns. From the plot shown above, we can appreciate a clustering of sorts in the left side of the plot which may indicate that the relationship between our variables are not linear in nature.

Given that we currently do not have to tools to address this, we will proceed with the assumption that the relationship between the dependent and independent variables are linear, and make note of this for further future analysis.

See Fig. (2).

4.3 Normality Assumption

For the model to be valid, the Normality test should be performed and the residuals must be normally distributed, the first thing that we did was a histogram to identify the condition, as we can see from the Figure, the normality seems to be really present there as there are no data points that seems to be at the end of the tails. At the same time we also plotted a normality plot for residuals which also seems to have a very normal condition with some outliers, but as the dataset is huge, outliers will loose importance.

$$H_0 : \text{The sample data is normally distributed}$$
$$H_a : \text{The sample data is NOT normally distributed}$$

For this test we tried to run the Shapiro-Wilk test but as the dataset consists of more than 5,000 the test was not able to run but as we have a huge number of data points we fail to reject the Null Hypothesis and conclude that the data used is indeed normally distributed.

See Figs. (3) and (4).

4.4 Equal Variance Assumption

$$H_0 : \text{Heteroscedasticity is not present}$$
$$H_a : \text{Heteroscedasticity is present}$$

The output displays the Breusch-Pagan test that results from the model. The p-value = 2.2e-16<0.05, indicating that we do reject the null hypothesis. For this reason, we will increase the power of title char count and the number of ratings. Using the poly function in R, we can test the higher order model to a power of 10 for these two predictors. Upon changing the powers of titlecharcount and number of ratings to the power of 10, the p-value = 2.2e-16<0.05. Again, we reject the null hypothesis in favour of the alternative hypothesis indicating that our model is heteroscedastic.

4.5 Multicolinearity Test

To look for multicollinearity in our final model we decided to look at multiple variance inflation factors (VIF) to help us best distinguish which variables may be highly correlated to each other. Our VIF values consisted of titleCharCount being 1.0681, averageRatingVotes to be 1.0077. For the categorical variables of our model, we only found the genre of Drama to detect multicollinearity. Since the only highly correlated variable in our data set seems to be a categorical variable, we ultimately cannot drop it and decided to keep it in our model.

We attempted to run ggpairs function to reinsure the highly correlated values were not present, unfortunately it was taking a substantial amount of time to give us any output (more than 3+ hours).

4.6 Influential Points and Outliers

We can see in the chart below that there might be some outliers present in the dataset, but as the data points are huge, we decided not to eliminate any of the outliers as they loose weight due to the amount of data present.

See Fig. (5).

5 Final Model

Based on the information provided above, we can conclude that our best fitted model is the reduced model which consists of: Title Char Count, Average Rating and Genre, and we can see it as:

$$\widehat{Rating} = \beta_0 + \beta_1 X_{title.char.count} + \beta_2 X_{Avg.Rating.Votes} + \beta_3 X_{genre}$$

Once we have the final model, we expanded to use the coefficients and get it as:

$$\widehat{Rating} = 4.972e^{-3} * title.char.count + 3.137e^{-3} Avg.Rating.Votes + \begin{cases} 6.162 - 0.6482 & = 5.5138 \longrightarrow Action \\ 6.162 - 0.3561 & = 5.8059 \longrightarrow Animation \\ 6.162 - 0.5152 & = 5.6468 \longrightarrow Comedy \\ 6.162 - 0.4283 & = 5.7337 \longrightarrow Crime \\ 6.162 + 0.8038 & = 6.9658 \longrightarrow Documentary \\ 6.162 + 0.02586 & = 6.1878 \longrightarrow Drama \\ 6.162 - 1.235 & = 4.9270 \longrightarrow Horror \\ 6.162 - 0.3262 & = 5.8358 \longrightarrow Romance \\ 6.162 + 0.6203 & = 6.7823 \longrightarrow Other \end{cases}$$

5.1 R^2_{adj} and RMSE of Best Fit Model

$R^2_{adj} = 0.1748$, this indicates that 17.48 percent of the variation of the response variable Ratings is explained by the best fit model with predictors Title Char Count, Average Rating Votes and Genre.

RMSE = 1.212 indicates that the standard deviation of the unexplained variation in estimation of the response variable Ratings is 1.212 rating.

5.2 Interpreting Coefficients

When holding all other variables constant, the estimated Rating will change by 4.972e-3 for each character in the movie title.

When holding all other variables constant, the estimated Rating will change by 3.137e-6 for each rating vote that the movie receives.

With regards to Genre, if we hold all other variables constant, the estimated Rating will change by the following amounts depending on the movies Genre:

- 5.5138 if the Genre is Action.
- 5.8059 if the Genre is Animation.
- 5.6468 if the Genre is Comedy.
- 5.7337 if the Genre is Crime.
- 6.9658 if the Genre is Documentary.
- 6.1879 if the Genre is Drama.
- 4.927 if the Genre is Horror.
- 5.8358 if the Genre is Romance.
- 6.7823 for all other Genre.

6 Discussion

The linear model was built to predict movie ratings based on the independent variables title.char.count, Avg.Rating.Votes and genre. To improve this model further, additional models were built that took into consideration interaction terms as well as higher orders of the predictors. It was determined that while these interaction terms and higher order terms were statistically significant (p-value < 0.05), the coefficient of determination of the model increased by such a small margin to where these improvements were deemed insignificant. For this reason, the final model does not include any interaction or higher order terms.

The model was tested against the key assumptions outlined in the methodology section of the report. These assumptions were multicollinearity, linearity, normality, homoscedasticity, outliers/leverage points and independence. The model that was generated validated the assumptions for multicollinearity and normality. However, the model did not satisfy the assumptions for homoscedasticity and linearity. Attempts were made to increase the order of the model to achieve homoscedasticity, but the p-value of the BP test remained significantly smaller than 0.05. Similarly, to achieve linearity, the model was transformed using the BoxCox transformation to find the best lambda value. However, similarly to the results for testing homoscedasticity, resulting residual vs. fitted plot for the transformed model still contained clustering and obvious trends, indicating that the model failed to satisfy the linearity assumption. We were not able to use the Shapiro-Wilks test to test for linearity because of the size of our dataset for the model.

The independence test provided our team with interesting results. There appeared to be a conical shape to the data as time progressed, which indicates that the model fails the independent test. However, in smaller chunks of time, there appears to be no trending of the error based on time. This indicates to our team that for future work, it might be advisable to break the data into smaller subsets based on the time. It is also important to note that a significant amount of data for this dataset was recorded in 2010 and onwards, compared to the rest of the data in the dataset.

For the reasons discussed above, the model that has been built is not suitable to make meaningful predictions for the rating of a movie. To try and built a suitable model, the next steps would be to return to the original dataset and determine if there are any other predictors that could be used in the prediction process that may have been missed. Other steps include looking for external data sources and combining additional data frames to increase the number of quality independent predictors which could be used in the model.

7 Figures and References

Grand View Research. 2022. “Movies and Entertainment Market Size, Share & Trends Analysis Report.” https://www.grandviewresearch.com/industry-analysis/movies-entertainment-market.

IMDb. 2022. “IMDb Datasets.” https://www.imdb.com/interfaces/.

Market Reports World. 2022. “Entertainment and Media Market Size & Share.” https://www.globenewswire.com/news-release/2022/02/11/2383700/0/en/Entertainment-and-Media-Market-Size-Share-2022-2028-

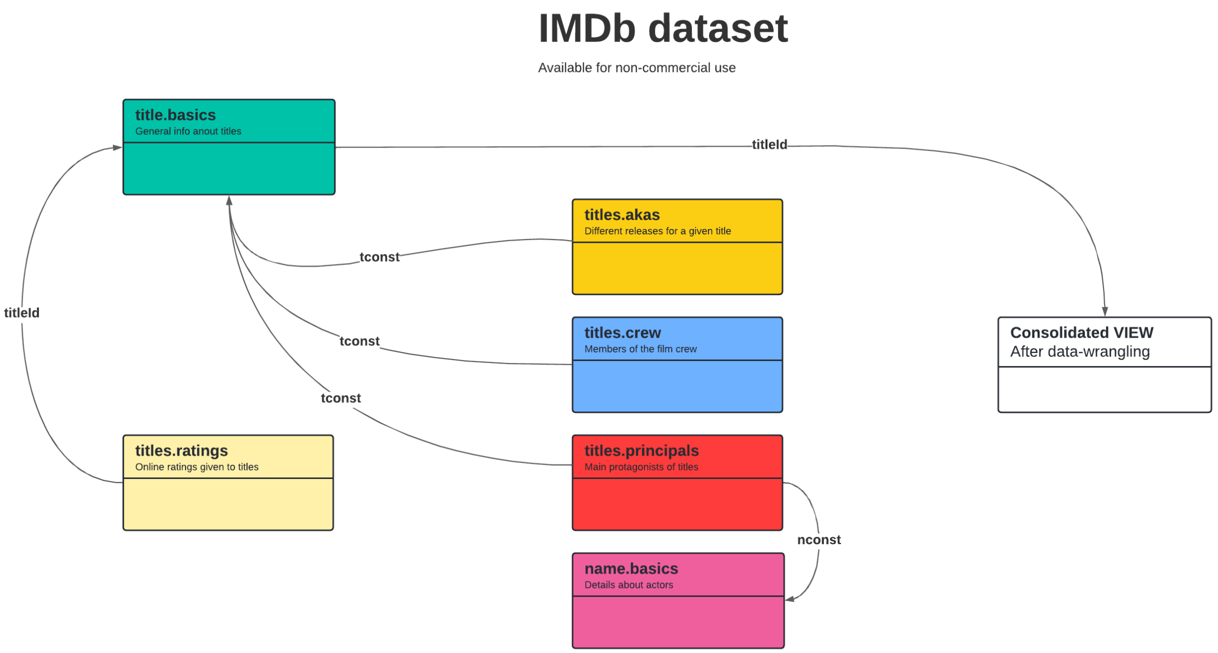

Figure 1: High level view of the dataset's internal structure. Row counts vary from tens of millions to a million or less depending on the table

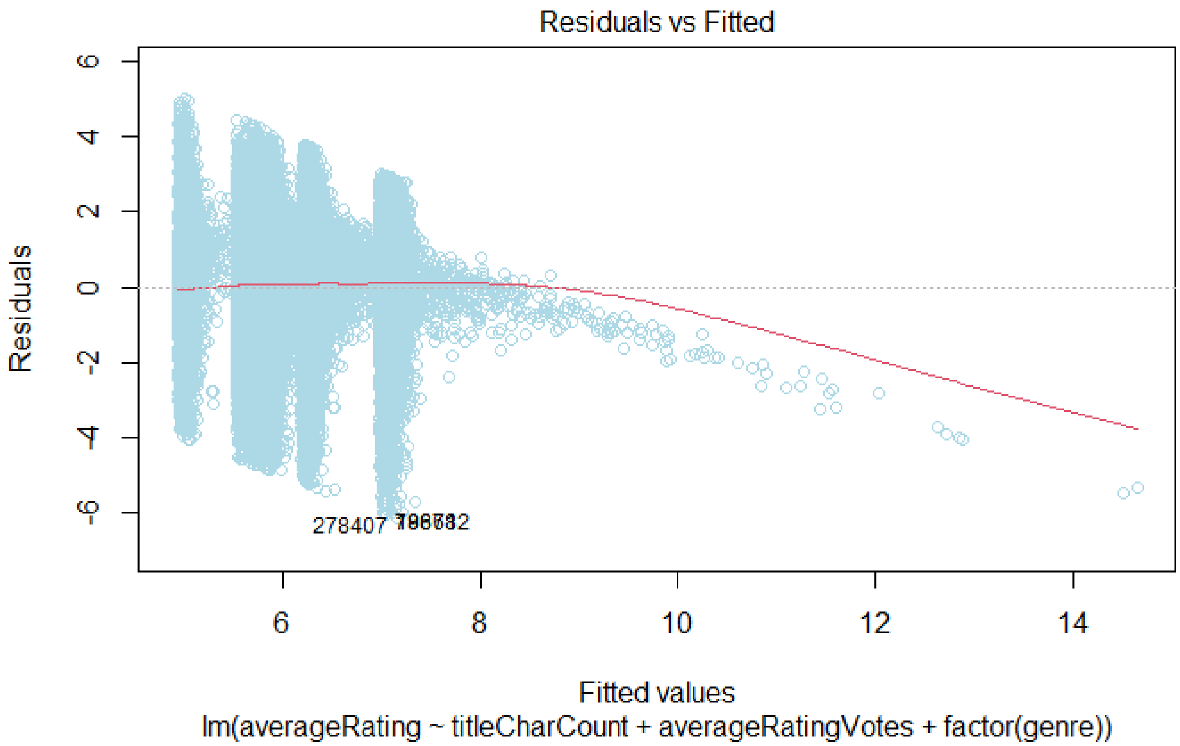

Figure 2: Linearity of residuals: heteroskedasticity present

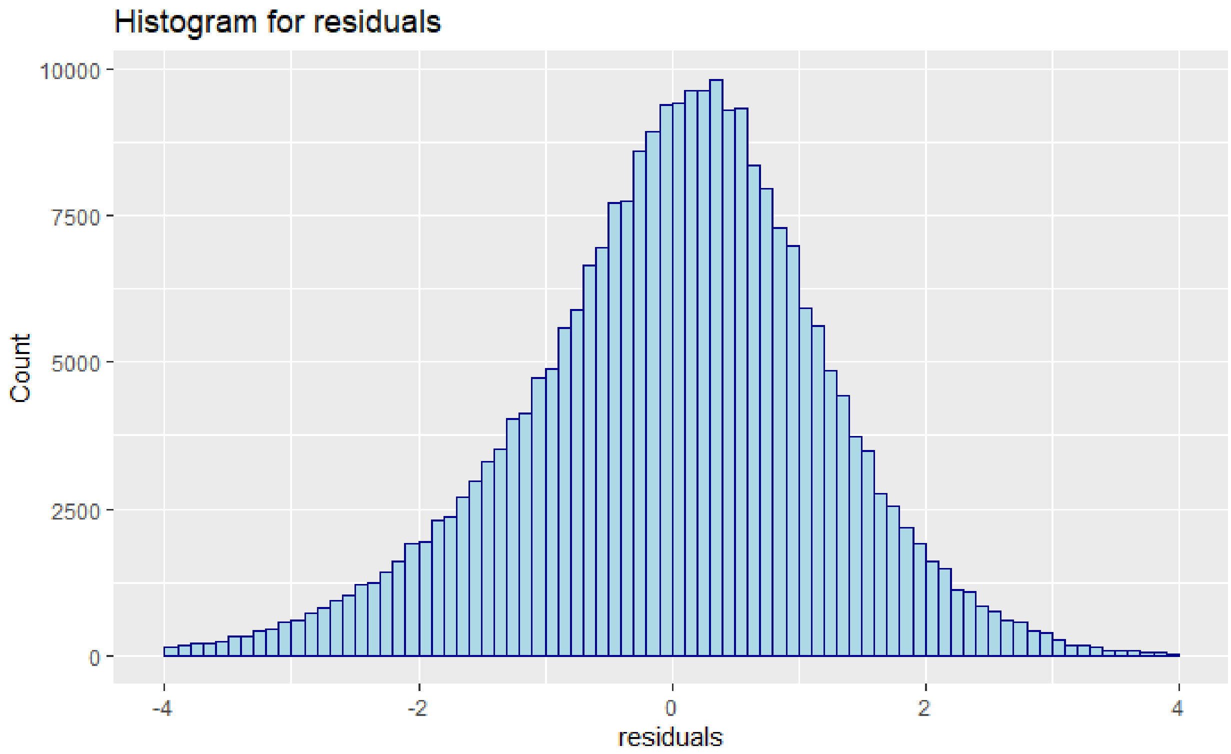

Figure 3: Normality of residuals

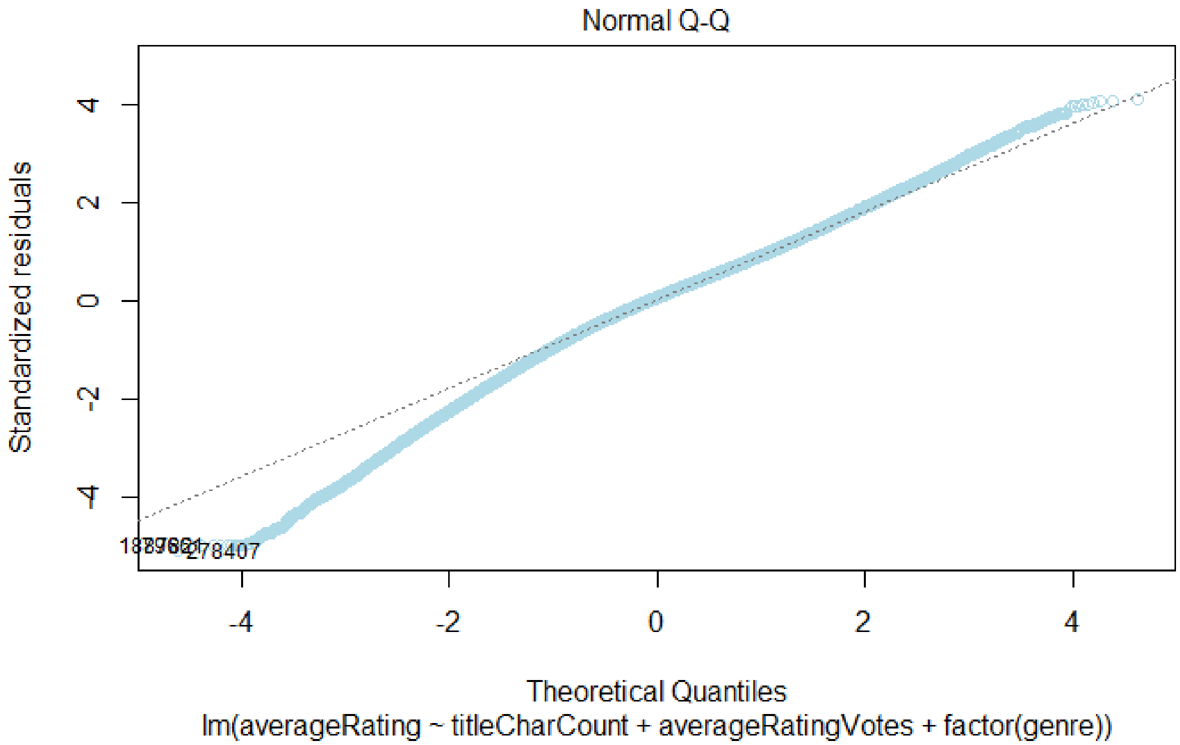

Figure 4: QQ Plot: normality of residuals

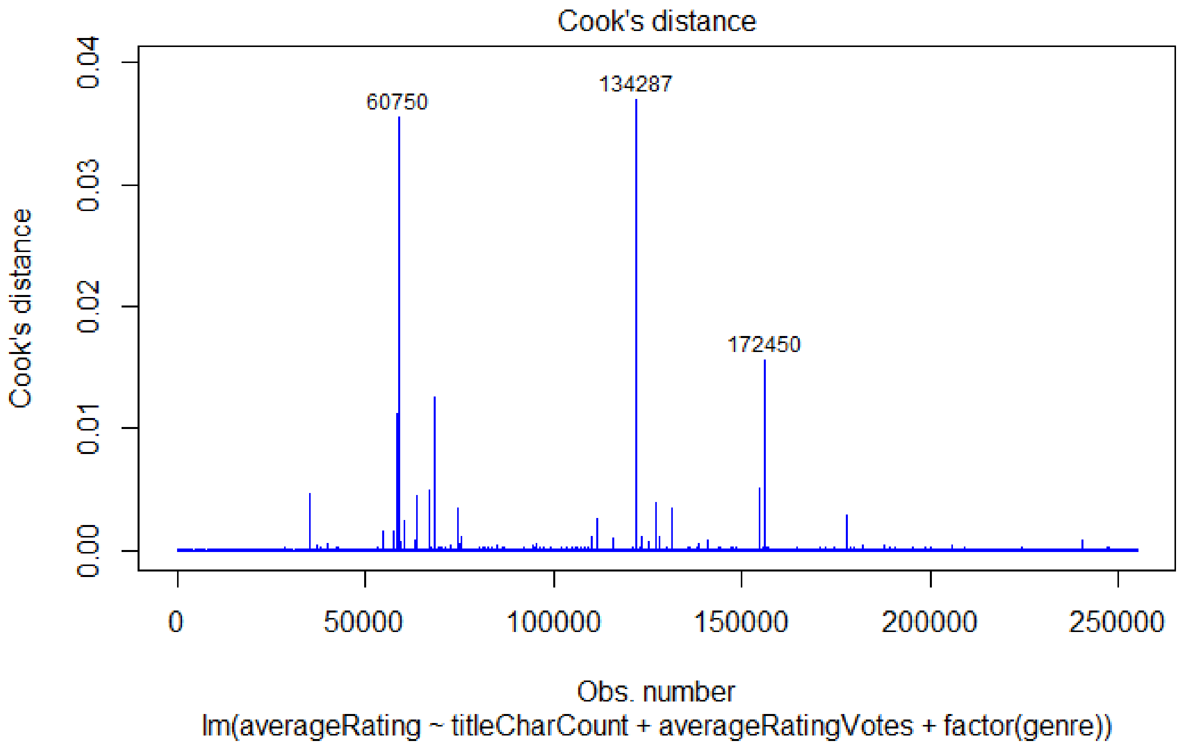

Figure 5: Cook's distance for outliers

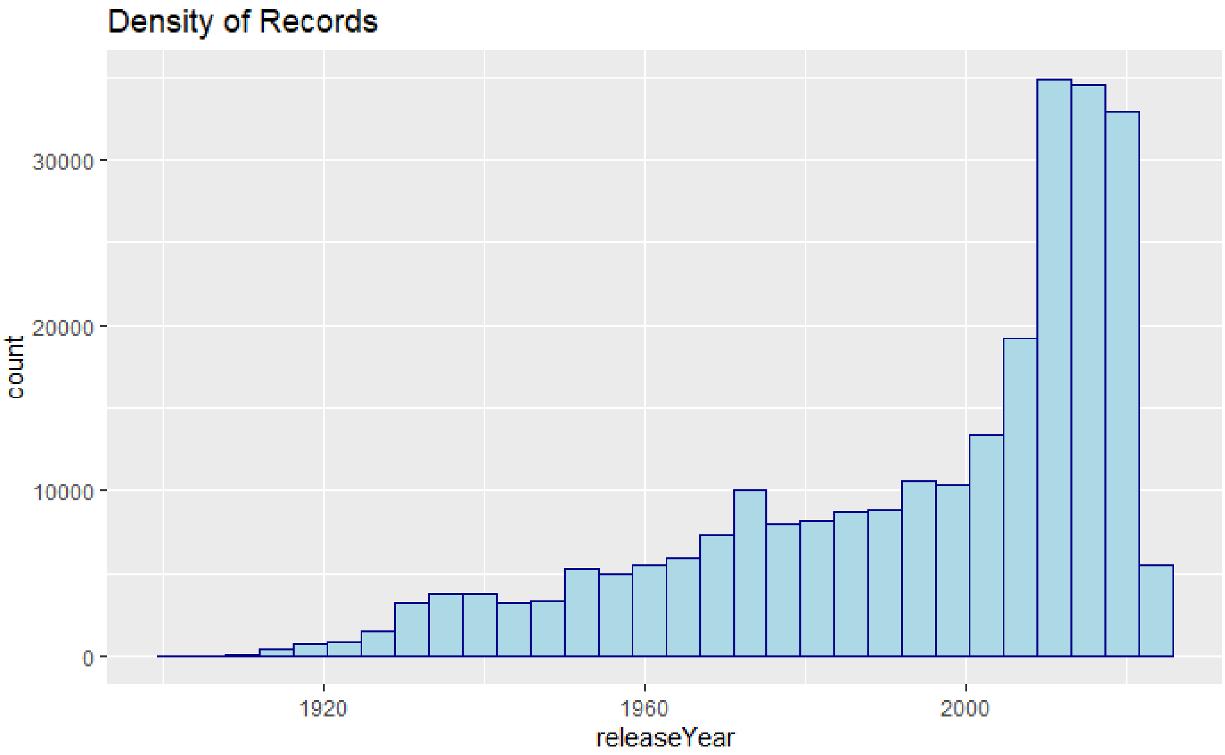

Figure 6: Density of records for entire dataset

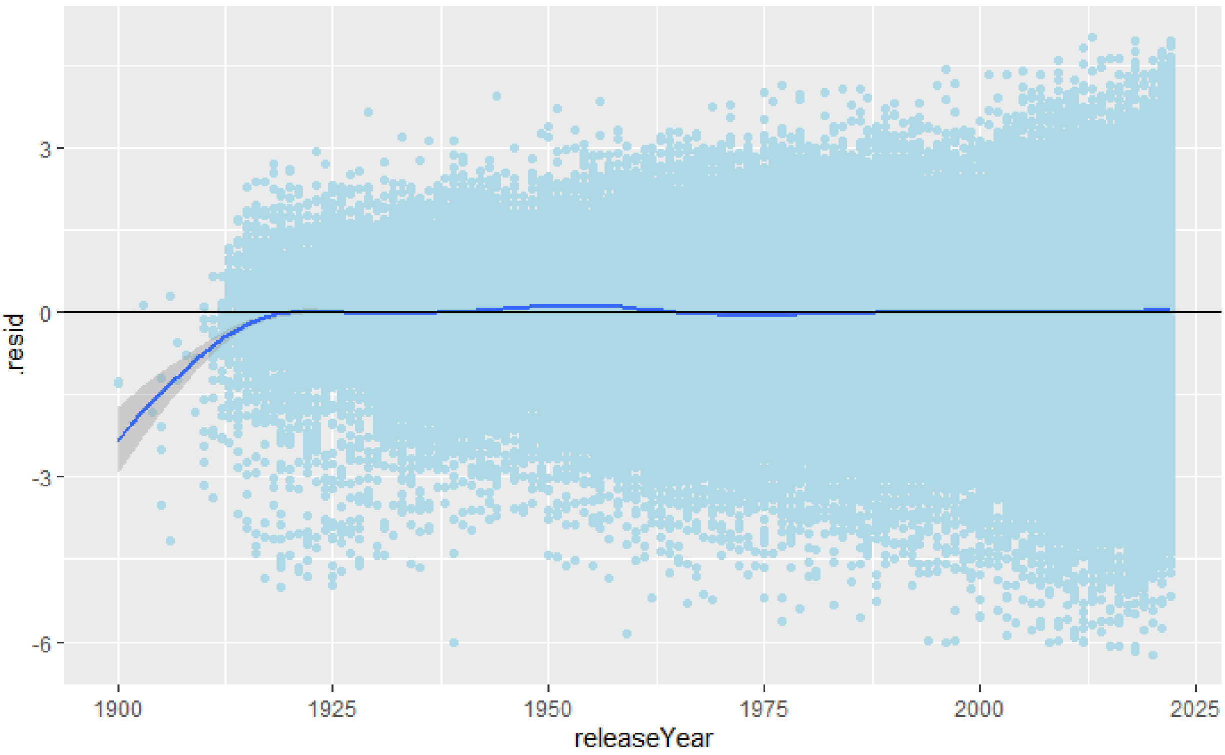

Figure 7: Independence

By-Industry-Demand-Global-Research-CAGR-of-5-9-Leading-Players-Market-Potential-Regional-Overview-Traders-Key-Findings-and-SWOT-.html.