How to make your movie a success

Executive Summary

This report showcases our analysis of a large film dataset and how key decision-makers can improve the odds of the popularity and the critical and commercial success of their films. Our key stakeholders are film production, distribution and marketing studios and firms.

To do this, our team acquired, prepared, and explored datasets from The Movie Database, gained from a free-to-use API. This included QQ plots and correlation analysis. We then analysed and modelled it using Linear Regression and a Generalized Linear Model. This helped describe and explain the dataset and the relations each factor had with one another.

The key factors that indicate film success are:

- Revenues
- Vote counts
- Average votes
- Popularity

Factors we analysed included:

- Genre, including whether the film was classed as ‘adult’
- Runtimes
- Budgets
- Starring actors.

This report also notes several limitations and potential issues to do with the report, its analysis and use. This Report makes several recommendations regarding:

- Film production
- Movie marketing
- Streaming and home video
- Further opportunities for exploration

Contents

Executive Summary 1

Rationale 4

Stakeholders 4

Research Questions 5

Acquiring and Understanding the Data 6

Datasets 6

Definitions 6

Limitations 7

Ethical issues 8

Data Preparation 9

Exploratory Data Analysis 9

Popular Movies 10

Profitable Movies 10

Movies with the Highest Budget 11

The Relative Frequencies of Genres 11

The Movies with the Longest Runtime 13

Vote Averages Across Genres 13

The QQ Plots 14

Correlation Analysis 15

Vote average and popularity 15

Vote average and budget 16

Modelling 16

Linear Regression 17

Generalized Linear Model (GLM) 20

Evaluation 22

Findings & Recommendations 25

Film Production 25

Film Marketing 25

Film Streaming 26

Further Opportunities 26

Conclusions 26

Appendix 27

1. References 27

2. Data Code Samples 28

Data acquisition processes 28

Data merger processes 29

Table of Figures

Figure 1: Missing Values 9

Figure 2: Top 20 Most Popular Movies 10

Figure 3: Top 20 Most Profitable Movies 10

Figure 4: Top 20 Movies with the Highest Budget 11

Figure 5: The Relative Frequencies of Different Genres 12

Figure 6: Top 20 Movies with the Longest Duration 13

Figure 7: Vote Averages Against Genres 13

Figure 8: The QQ Plot 14

Figure 9: Correlation Plot 15

Figure 10: Correlation Analysis 15

Figure 11: Popularity against Vote Average 16

Figure 12: Budget against Vote Average 16

Figure 13: Linear Regression Model with 5 Significant Input Variables 17

Figure 14: Residuals versus fits plot 18

Figure 15: Regression Plot of Movie Popularity vs Vote_Average 19

Figure 16: Regression Plot of Movie Budget vs Vote_Average 19

Figure 17: Regression Plot of Movie Revenue vs Vote_Average 20

Figure 18: Regression Plot of Movie Runtime vs Vote_Average 20

Figure 19: GLM 1 with 5 significant Input Variables 21

Figure 20: GLM 2 with Genre Type Input Variables 21

Figure 21: Evaluation Measure of Linear Regression & GLM 22

Figure 22: New Linear Regression Model with Vote_Rate 23

Figure 23: Confusion Matrix of the New Linear Regression Model 23

Figure 24: Evaluation Measure of the New Linear Regression Model 23

Figure 25: New Generalized Linear Regression Model 24

Figure 26: Confusion Matrix of the New Generalized Linear Regression Model 24

Figure 27: Evaluation Measure of the New Generalized Linear Regression Model 24

Rationale

Analyses of a film's success are typically performed through qualitative and creative lenses, however as commercial productions they are also a trove of potential data driven insights. Films are defined not only by their screenplay or reception at premiere, but by factors such as budget, casting choices, running time, public interest and long-tail performance. These datapoints are readily accessible to the public and present an opportunity to examine and produce films using a statistically driven approach.

There is precedent in such approaches used in the sporting industry, particularly that of baseball sabermetrics (Marchi 2018, ch4.4), also popularised as 'Moneyball' by Michael Lewis (Lewis 2003). Sabermetrics is the practice of representing baseball in only its numerical qualities to determine managerial choices. By putting ignoring traditionally 'obvious' qualitative indicators of success, potentially hidden insights can be gained.

In the film industry, this approach would resemble setting aside the quality of movie's story, its visual effects or its soundtrack. It would pay less heed to film festival receptions and publication reviews which place immense weighting to a handful of individual critics. The intent of this approach is not to replace the creative decision process, but to procure insights that focus groups and user studies would normally be relied upon for, but it is a less noisy method of gaining insights for stakeholders.

This will also make optimising film production and marketing along particular variates more feasible. As an example, beyond optimising a film for only critical reception and box office success, it could also attempt to maximise long-tail success on streaming services, home video sales and cultural significance. For already-released films, streaming and sales vendors could adjust their distribution to capitalise on predicted spikes in user interest.

Movies captivate us with their stories, but their statistical data has a story to tell as well.

Stakeholders

Our external stakeholders and target audience are movie production studios, particularly those with the appetite or means to follow data-driven decisions. They include long running companies such as Paramount, Universal Pictures and Village Roadshow Pictures as well as newer ventures in the industry like Netflix, Amazon and Hulu that have begun producing their own in-house movies.

The latter variety are perhaps most receptive to a 'Moneyball'-esque approach, given they are less attached to traditional Hollywood actors, production methods and distribution. As of 4th September 2021, Netflix and Amazon have respectively increased their number of original releases by 27% (Netflix 2021) and 55% (Amazon 2021) year-on-year compared to the same period in 2020. Their expansion into the movie production industry indicates both strong demand and resources. Additionally, by virtue of being online user platforms, they also possess a wealth of readily-accessible user data to leverage.

Smaller and independent studios would also be well-suited to using this approach as they have less inertia from production changes, though their capacity to action data insights may be lower than major studios.

While the examples given thus far are Western studios, the target audience is international. Our chosen dataset includes observations from many countries and the reception to foreign films by Western audiences is only increasing (Oscars 2021).

Internally, we have data scientists, statisticians, programmers and analysts involved in the production of this plan, comprising of:

- Jean Koh
- Jordan Daly
- Maria Dhaliwal
- Maggie Chen
- Rashmi Raveena
- Ron Au

Our key intent is to identify insights and assess their real world actionability and weaknesses.

Research Questions

The burning question that our analysis seeks to address is "How can production studios and streaming platforms improve the performance and popularity of their movies?" Stemming from this, several sub-questions arose that are integral to ensuring the success of movies:

- Is there a relationship between a starring actor's popularity and movie ratings?
- Do certain genres generally perform better or worse than other genres?
- Do films perform better if they're of a certain runtime?
- Does increasing a film's budget improve its critical performance?
- Is there still a relationship between actors, genres, runtimes and movie performance when controlling for a budget?
- Does the death of a starring actor improve a movie's ratings?

Based on the sub-questions stated above, we can see the occurrence of some factors to look into with our exploratory data analysis and regression models. We will investigate all these factors and utilise whatever data we can get our hands on to provide some recommendations.

Acquiring and Understanding the Data

Datasets

Our team considered using several datasets for our analysis. After numerous discussions, we managed to narrow the choices down to two viable datasets; IMDb and TMDb.

The Internet Movie Database (IMDb) provides a variety of downloadable datasets via their website (IMDB 2021), and also has a few APIs. Although the datasets provide a range of information on a variety of films, their use would not adequately demonstrate our data acquisition skills. With regards to the APIs, there is an API that is mainly for enterprise use and would demonstrate our data acquisition skills, it is not easily accessible. We have submitted an access request, but this is yet to be granted. The other APIs for consumer use are just sample datasets and would not be suitable for our research.

An alternative to the IMDB datasets and APIs is The Movie Database's (TMDB's) API (TMDB 2021). Their API is free to use and provides the level of detail required for our analysis on the research questions. As such, we elected to use it.

Definitions

Our TMDB dataset has the following columns defined in Table 1, below:

Table 1 Dataset Terms

ID	Movie Identification Number
Title	Movie's Title
Original_title	Movie's Original Title
Overview	Brief plot of the movie
Release_date	Date of release
Original_language	Movie's Original Language
Genre_ids	Movie's genres in categorical values
Adult	A logical value that indicates if a movie is only for adult people (TRUE) or not (FALSE)
Popularity	An actor or film's popularity on TMDB. This is used as a proxy for the personal popularity of each, but an actor's or film's popularity on the website may not reflect their drawing power in theatres
Vote_count	The number of votes

Vote_average	The average of the votes
Tagline	The movie's tagline
Runtime	Total duration of the movie
Budget	Gross budget in US dollars
Revenue	The movie's income in US dollars

Our team figured that the genre_ids column will not be very useful because the genres are in categorical values. We subsequently extracted the movie's genres in Boolean values, which helps to easily identify movies with multiple genres.

Besides our dataset, we have also clarified some terms under the research questions.

- Starring is the billing order in the movie.
- Death of an actor is the date of when a particular actor has died.
- Ratings and performance are the mean ratings for the movie, which is provided by the TMDB users.

Limitations

Our team does not have access to primary data. For example, TMDB only provides access to gross budget and revenues rather than other financial indicators such as net profit or Earnings Before Interest, Taxes, Depreciation, and Amortisation due in part to 'Hollywood accounting' (Sparviero 2015, p20). This has limited the scope of our research and the applicability of our claims and recommendations. To account for this, we have emphasised correlations and explanations rather than outright prediction. Budgets and revenues are listed in United States Dollars on TMDB and are not adjusted for inflation, exchange rates or purchasing power parity. This makes comparison between films easier but also may over-or-under-estimate the profitability of movies made outside of the United States.

Additionally, our team may not be able to gain full insight from the datasets as we are not all well-versed in every potentially-relevant regression technique. We have accommodated for this by sharing information on Microsoft Teams and analysis on GitHub. In this, we also aim to contribute to the wider data science community by making this information publicly-available.

The information available from TMDB is user-sourced so it varies from other websites such as IMDB. For example, some films reviews can have skewed results due to 'review bombing', which is when users coordinate to give certain reviews *en masse* (Tomaselli et al 2020). This may apply to TMDB. Information on TMDB's userbase is limited, which creates a risk of biased data. Our team has accounted for this risk. For example, we have eliminated clear outliers and biased votes by setting minimum requirements for film popularity and vote counts when selecting films for analysis. This reduced our sample size from around 9000 films to around 200. Websites that use a 5-point basis for ratings are generally consistent with one another (Gupta et al 2010, p37) and this is true of TMDB. So, after trimming down potentially-skewed ratings, we can make the data more consistent and useful as the TMDB ratings can be generally assumed to reflect the quality of the films themselves. While we are working with a smaller sample size, the data are of a higher quality.

The largest limitation of our data analysis is if it is underutilised or not fully implemented. For example, if our team's recommendations are not properly employed by production and marketing companies, this reduces their usefulness.

Ethical issues

There are many ethical issues to do with data analysis and the use of that data.

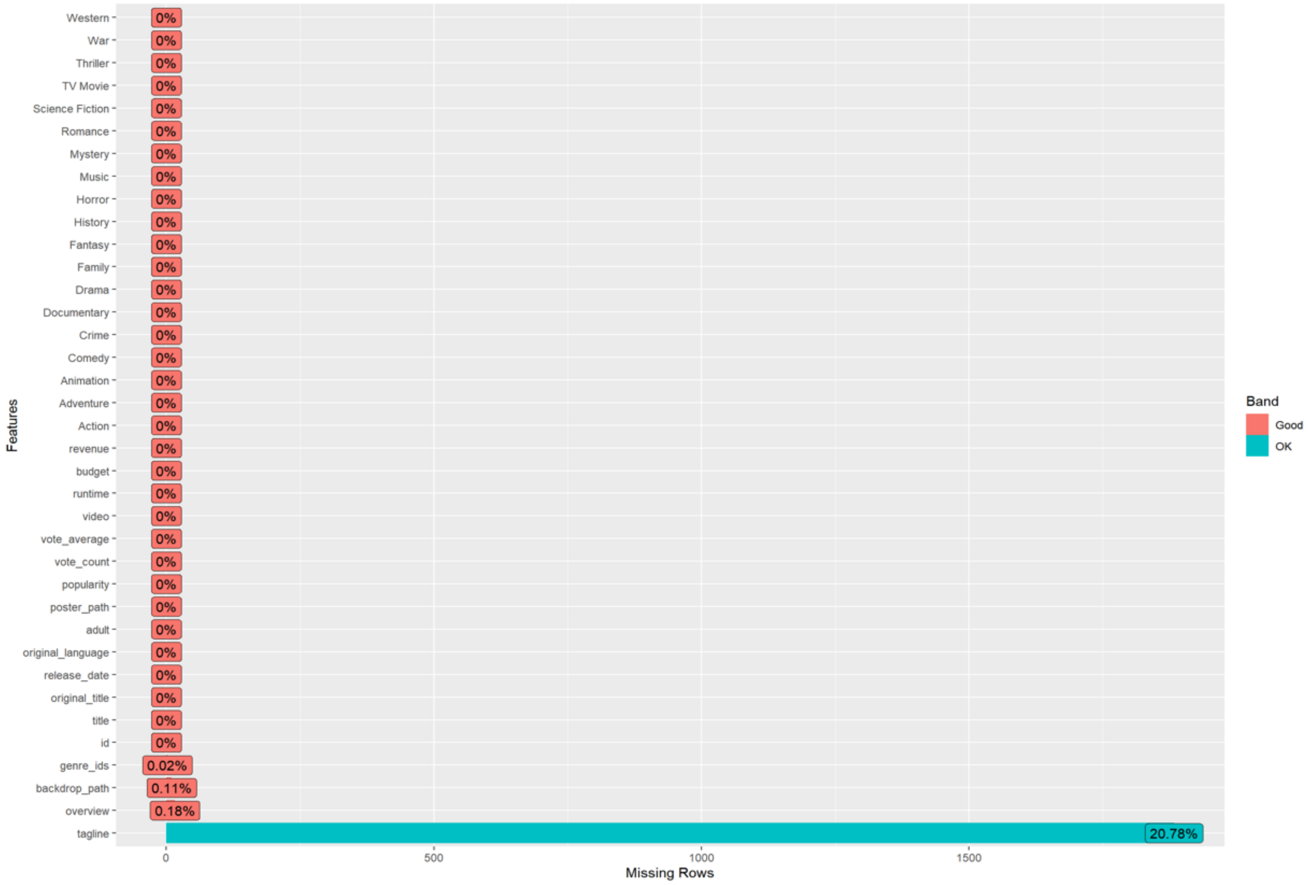
Privacy and consent when collecting and disseminating user data is a paramount concern for our team. This is adequately addressed as permission has been granted and for the use of the datasets and ownership of the original data has been acknowledged (Tripathy 2013, p1478) by this report. This is because we are using anonymised data collected from a third-party site which ensures confidentiality and consent.

Beyond privacy and consent, the most pertinent of the ethical issues to do with the implementation of data analysis is data slippage (Markham 2018, p1), where data are presented out-of-context and may be used to obfuscate or justify unethical decision-making (Breidbach 2020, p171). By presenting our data within context and with caveats, we hope to avoid its unethical use. For example, we find that an actor's death had a positive impact on movie popularity and ratings, as it may have for Heath Ledger's death before the release of The Dark Knight (2008). In this case, an unethical use of this data would be to kill actors to encourage higher engagement with a given film. However, we recommend against this as it runs counter to the intent of our analysis, especially given it describes and explains relationships rather than predicts trends. Our recommendations must be read in context in order to ensure they are employed ethically.

Data Preparation

Our team took the following steps to pre-process the data prior to the exploration of the data. Firstly, we have checked if there are any missing values in the dataset. Figure 1 shows there are some missing values in the dataset. Those variables are genre_id, backdrop path, overview and tagline. As this is a huge dataset, we have decided to drop the columns with the missing values.

Figure 1: Missing Values



Exploratory Data Analysis

Before conducting regression analysis, exploratory data analysis is an approach of analysing datasets to summarise the main characteristics and to spot any anomalies. This is done with the help of summary statistics and graphical representations. Please note that the statements below are regarding the TMDB dataset, not a dataset of all films made.

Popular Movies

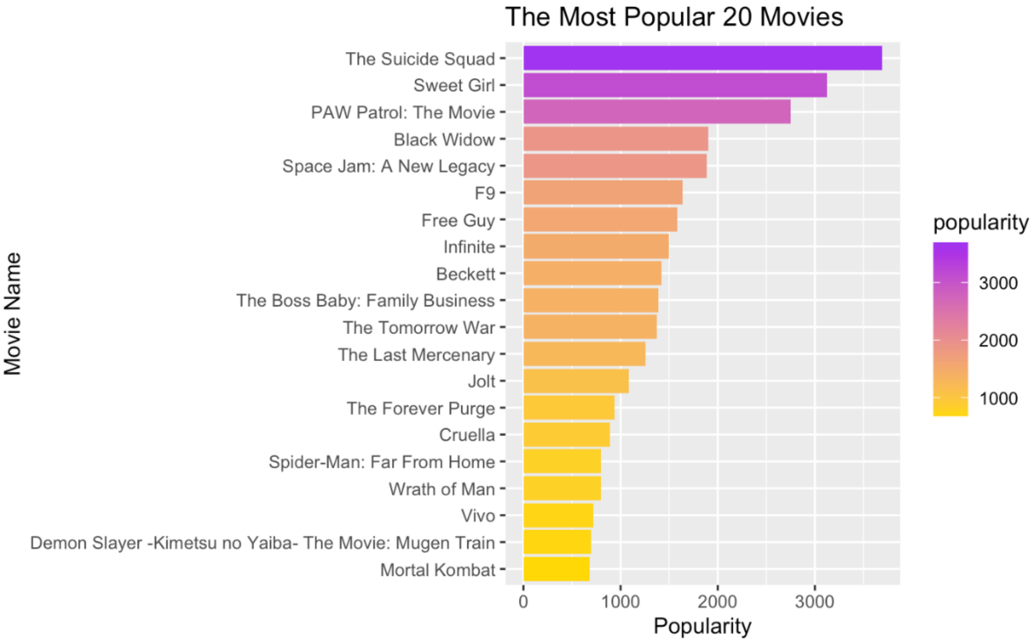


Figure 2: Top 20 Most Popular Movies

Figure 2 shows the movies with the highest popularity, and we have selected the first 20 popular movies. It shows that "The Suicide Squad" is the most popular movie.

Profitable Movies

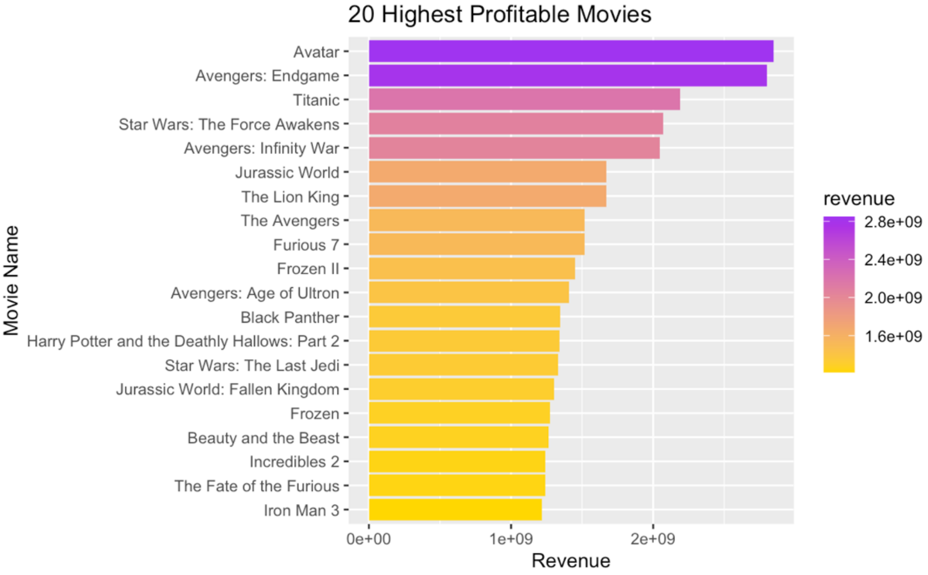


Figure 3: Top 20 Most Profitable Movies

Figure 3 shows the top 20 movies that have generated the highest revenue. It can be seen that "Avatar" and "Avengers' Endgame" have significantly generated more revenue compared to other movies.

Movies with the Highest Budget

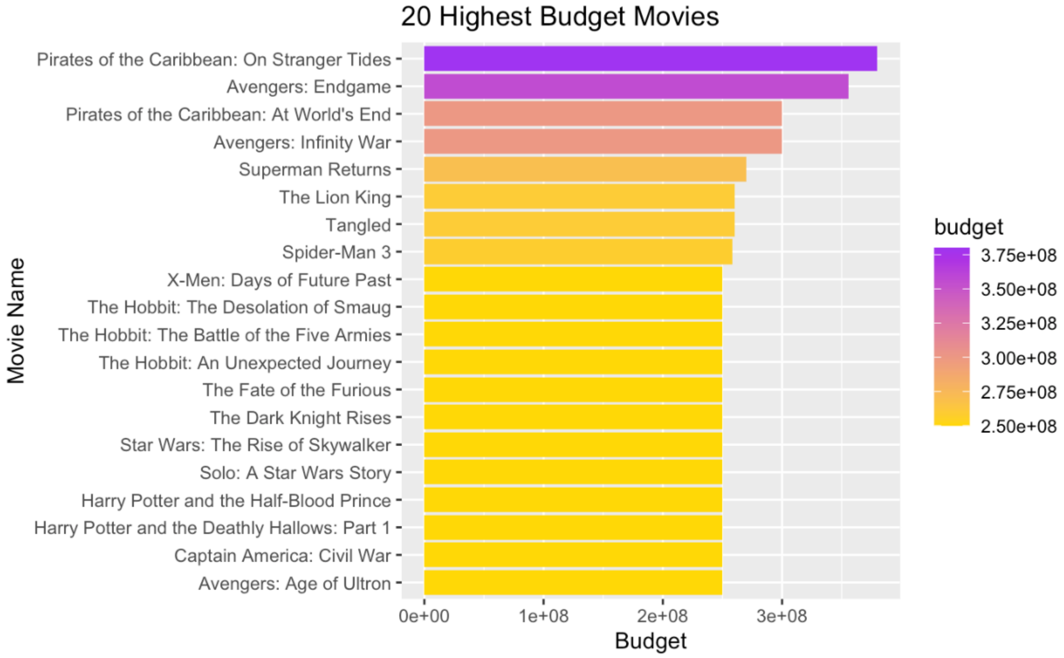
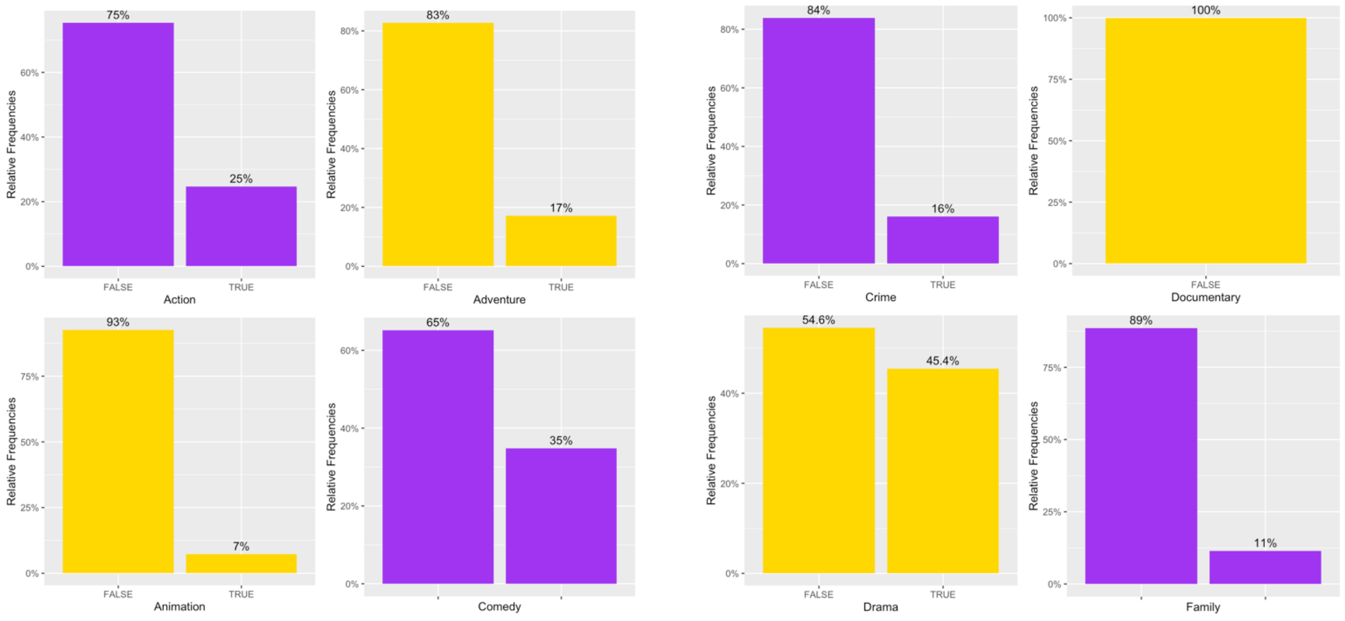


Figure 4: Top 20 Movies with the Highest Budget

Figure 4 shows the top 20 movies with the highest budget and "Pirates of the Caribbean: On Stranger Tides" is the highest budget movie.

The Relative Frequencies of Genres



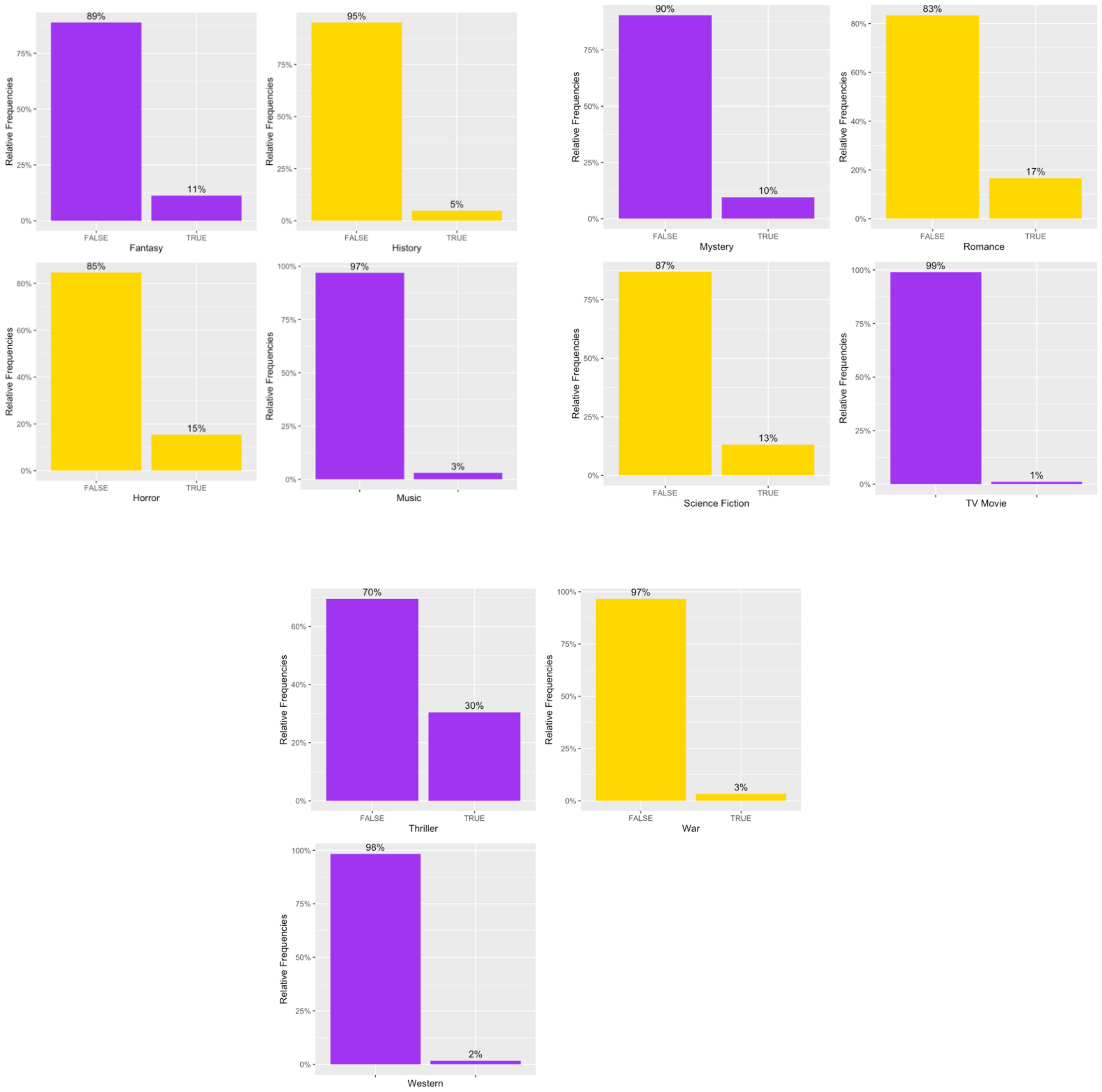


Figure 5: The Relative Frequencies of Different Genres

The relative frequencies of the kinds of genres are also analysed. It can be seen that the most common genres in our dataset are Action, Comedy, Drama and Thriller. Some other genres are uncommon, especially Documentary and TV Movie. This particular information indicates that none of our movies in our dataset is a documentary or TV Movie.

The Movies with the Longest Runtime

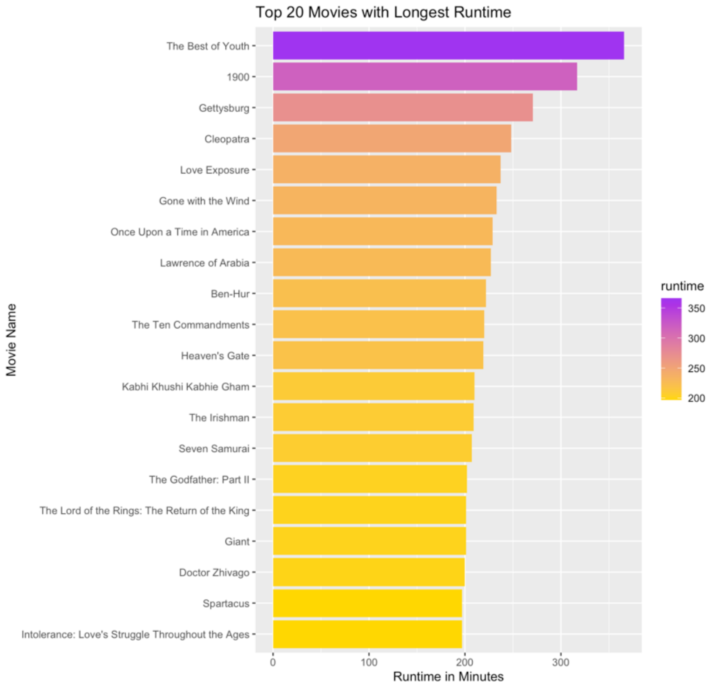


Figure 6: Top 20 Movies with the Longest Duration

A simple bar chart in Figure 6 is also generated to show which movies have the longest duration.

Vote Averages Across Genres

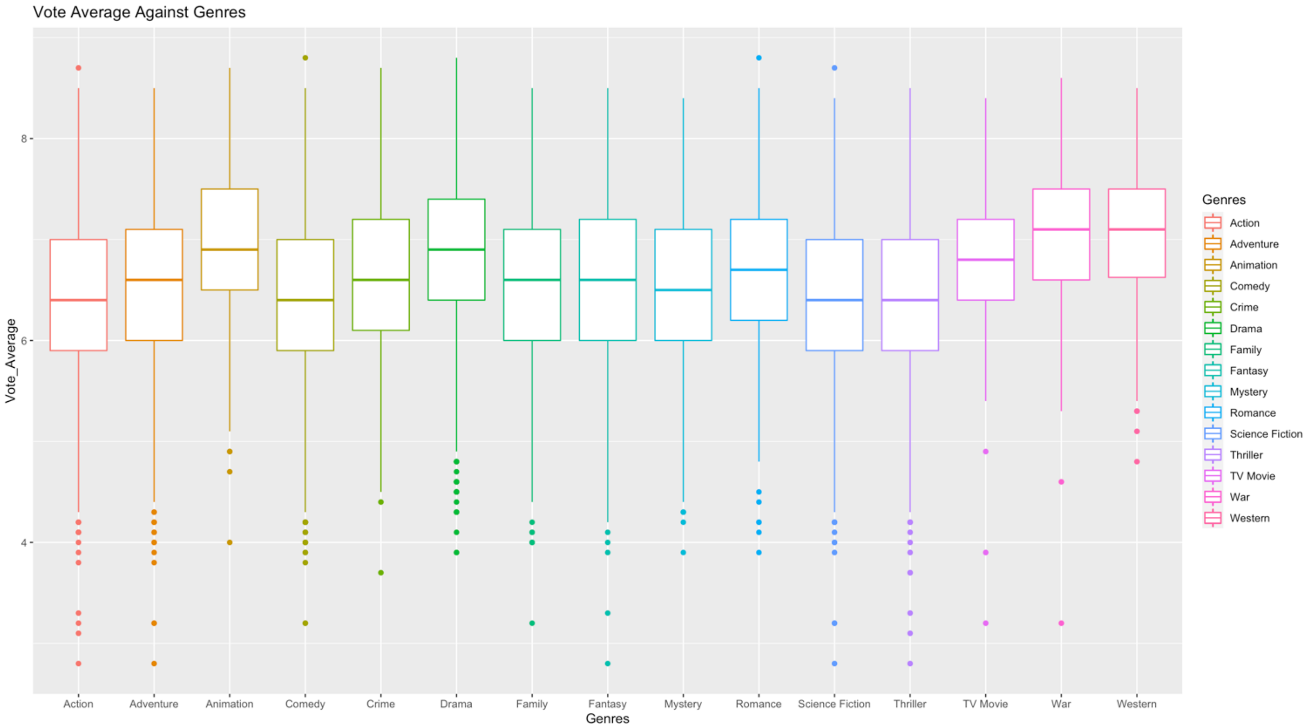


Figure 7: Vote Averages Against Genres

The vote averages are analysed against the different kinds of genres. It can be seen that the median value for most of the genres are between the vote average, 6 and 7. Some genres also have more outliers, compared to other genres and the outliers show that some films are rated way lower than majority of the films.

The QQ Plots

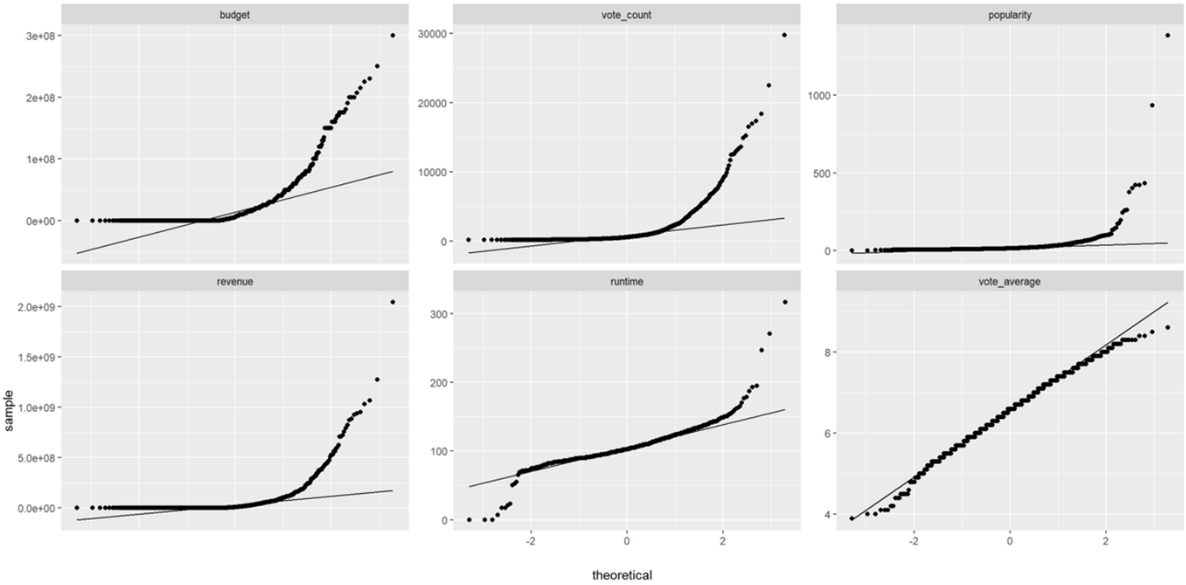


Figure 8: The QQ Plot

The QQ plot shows whether the set of data came from some theoretical distribution such as a normal or exponential. If the data points are fitted into a straight line, it shows the data are normally distributed. If the data points are not fitting into a straight line, it shows that the data are distributed as the standard normal. According to Figure 8, it shows that some data points are fitted into the line. So, the linearity of the points suggests that the data are normally distributed in these plots.

Correlation Analysis

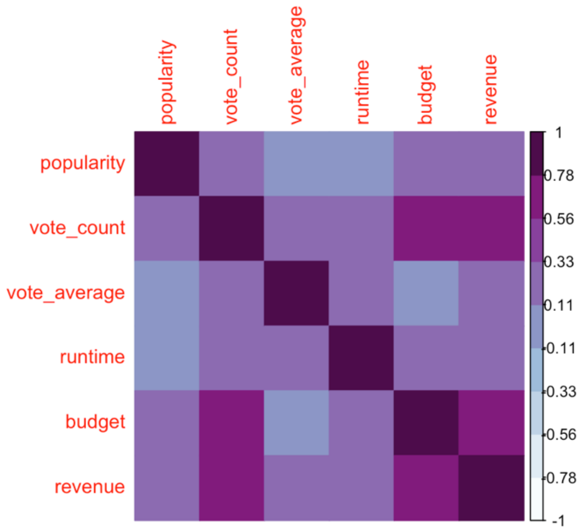
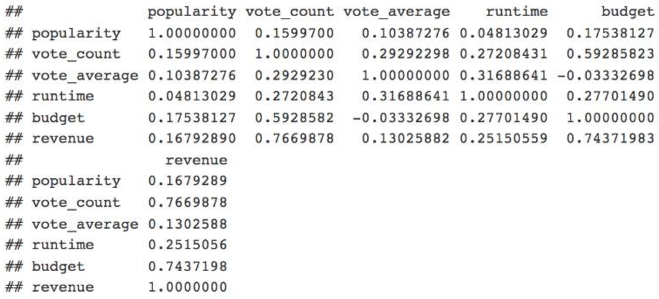


Figure 9: Correlation Plot



```
##              popularity vote_count vote_average    runtime      budget
## popularity   1.00000000  0.1599700   0.10387276 0.04813029  0.17538127
## vote_count   0.15997000  1.0000000   0.29292298 0.27208431  0.59285823
## vote_average 0.10387276  0.2929230   1.00000000 0.31688641 -0.03332698
## runtime      0.04813029  0.2720843   0.31688641 1.00000000  0.27701490
## budget       0.17538127  0.5928582  -0.03332698 0.27701490  1.00000000
## revenue      0.16792890  0.7669878   0.13025882 0.25150559  0.74371983
##                revenue
## popularity   0.1679289
## vote_count   0.7669878
## vote_average 0.1302588
## runtime      0.2515056
## budget       0.7437198
## revenue      1.0000000
```

Figure 10: Correlation Analysis

Correlation analysis is useful to check whether there are possible connections between variables. In this case, we have specifically selected the following variables:

Vote average and popularity

The voting average and popularity show a positive correlation value of 0.10387276. This means that the two variables move in the same direction and they have a strong relationship. Figure 11 shows clearly the positive relationship.

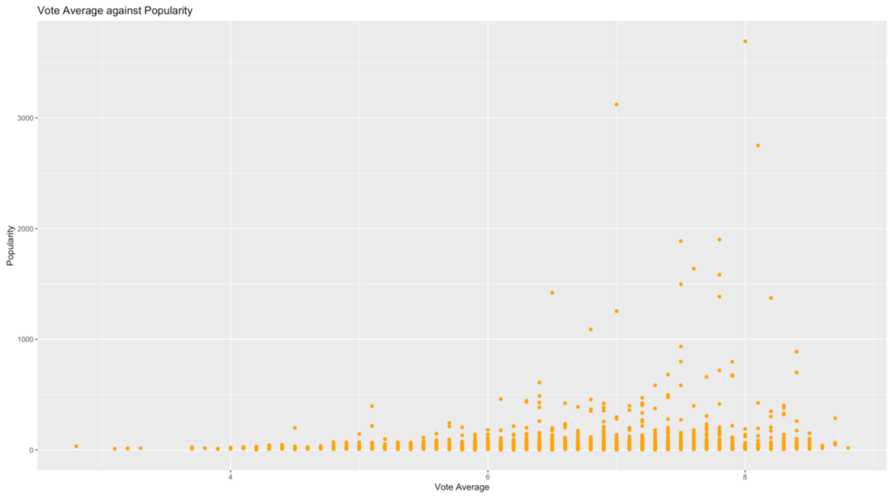


Figure 11: Popularity against Vote Average

Vote average and budget

The vote average and budget variables show a negative relationship of -0.03332698. This is an interesting point which shows that voting average will not depend on the variable budget and Figure 12 shows their weak relationship.

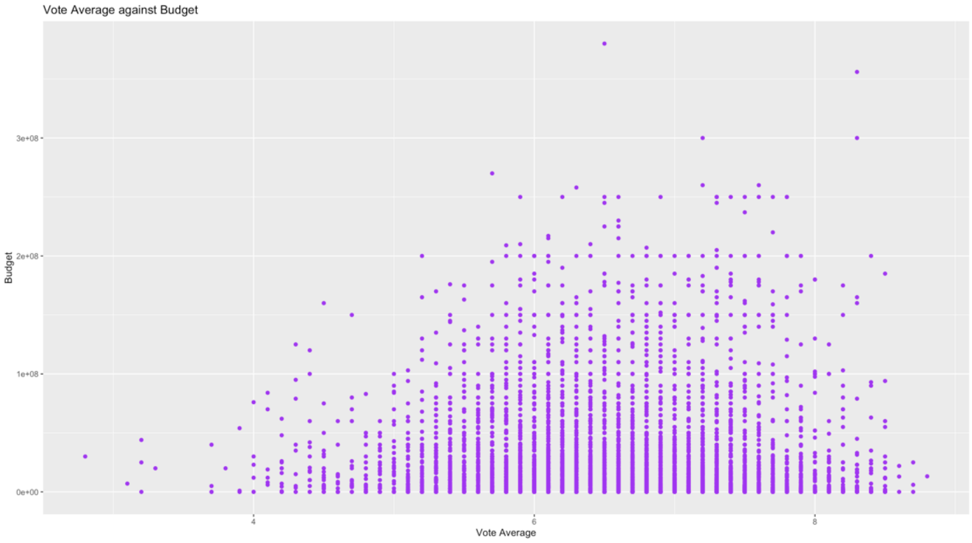


Figure 12: Budget against Vote Average

Likewise, we can get a clear understanding about the relationship of variables by checking the Figures 9 and 10 above. Overall, the positive values show that there is a strong relationship between the variables and the negative values shows that there is weak relationship between variables.

Modelling

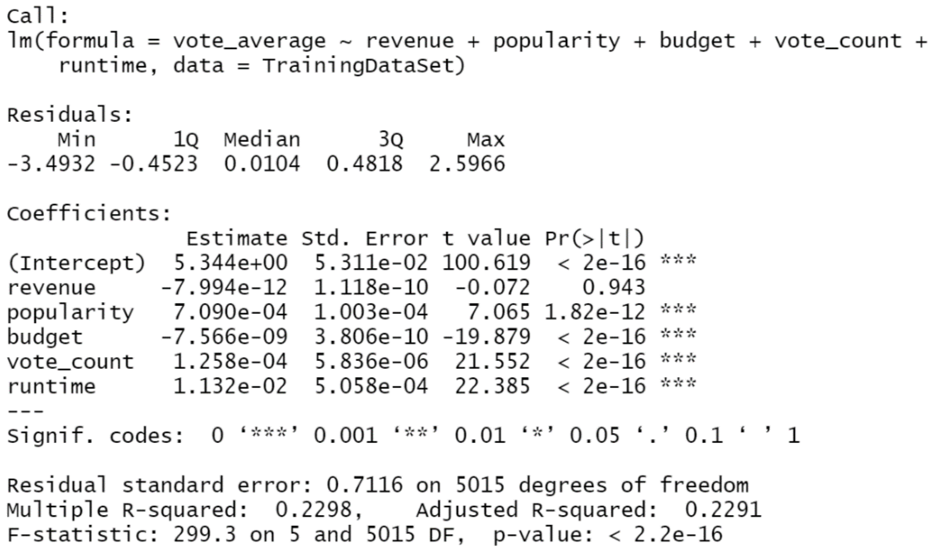
Our modelling observed the relationship among variables by using linear regression model. After data preparation, we selected five significant input variables (x) to study the linear relationship with the output variable vote_average (movie rating). These are:

- Revenue,
- Popularity,
- Budget,
- Vote count, and
- Runtime.

The training sample is used to conduct modelling, while the testing sample is used to test the prediction's accuracy. It can create multiple prediction models by allocating varying proportions of training and testing samples.

Linear Regression

For the below linear regression model, we took 70% of the data as the training dataset and the remaining proportion as testing dataset. It demonstrated that there is close to 71% of the variability in audience score can be explained by our model. The p-value is lower than 5%, indicating that there is a more than 95% probability that the model is correctly stated. All the variables analysed have positive correlation with rating, while budget has negative correlation with movie rating. It depicts that even though a movie has a higher budget, it does not mean that it will be successful.



```
Call:
lm(formula = vote_average ~ revenue + popularity + budget + vote_count +
    runtime, data = TrainingDataSet)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4932 -0.4523  0.0104  0.4818  2.5966

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.344e+00  5.311e-02 100.619  < 2e-16 ***
revenue     -7.994e-12  1.118e-10  -0.072    0.943
popularity   7.090e-04  1.003e-04   7.065 1.82e-12 ***
budget      -7.566e-09  3.806e-10 -19.879  < 2e-16 ***
vote_count   1.258e-04  5.836e-06  21.552  < 2e-16 ***
runtime      1.132e-02  5.058e-04  22.385  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7116 on 5015 degrees of freedom
Multiple R-squared:  0.2298,	Adjusted R-squared:  0.2291
F-statistic: 299.3 on 5 and 5015 DF,  p-value: < 2.2e-16
```

Figure 13: Linear Regression Model with 5 Significant Input Variables

From the residual versus fits plot, it indicated that the residuals bounce around the 0 line at random, which implies that the assumption of the linear relationship is reasonable. And around the 0 line, the residuals form a horizontal band, which determines that the variances of the error terms are equal.

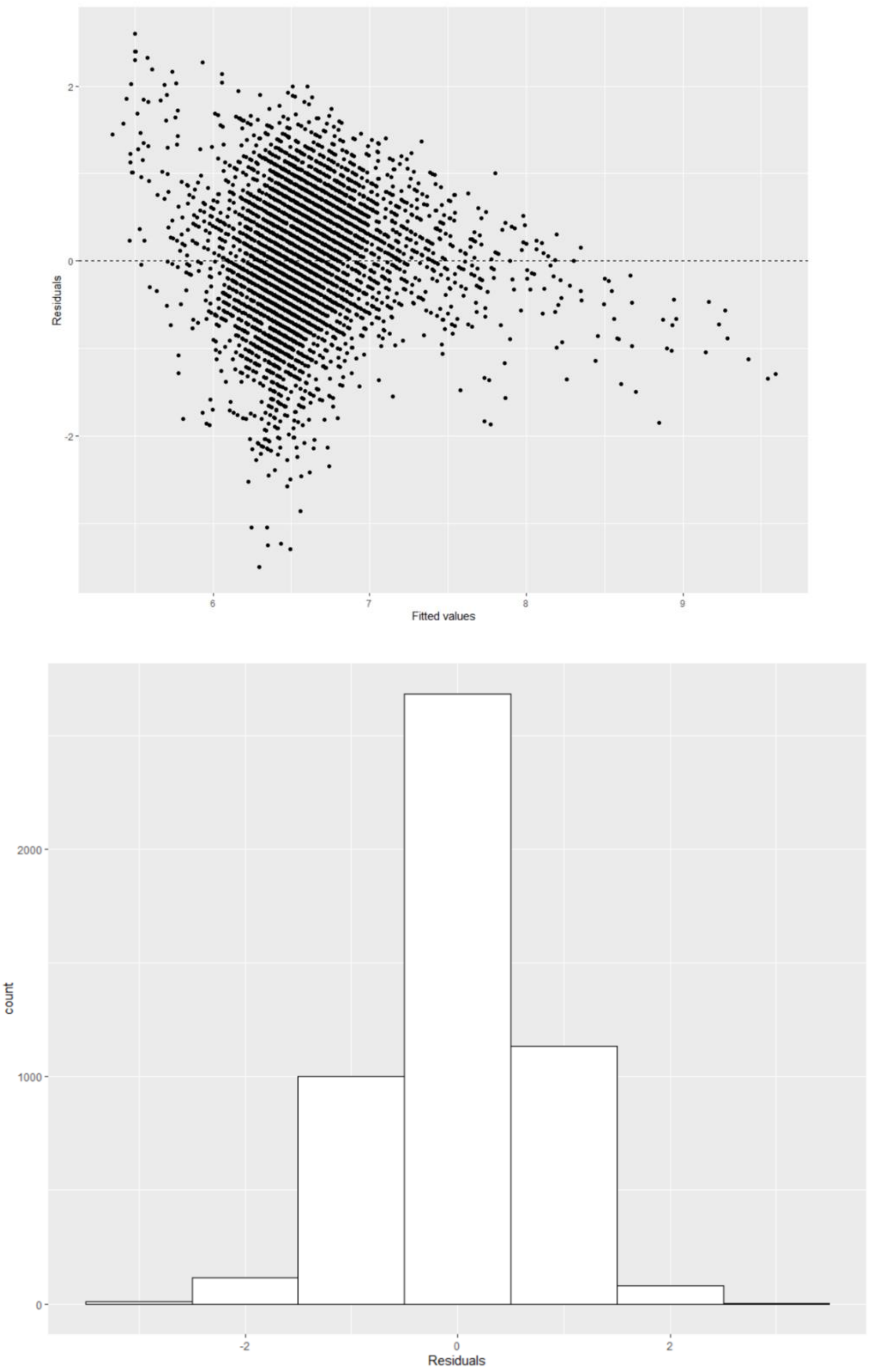


Figure 14: Residuals versus fits plot

Then we visualised the model and generated some linear regression plots between each individual input variables and the output variable.

The first plot shows a clear trend of higher rating when popularity is higher. That means when the casts of the movie or the movie itself are popular on TMDB, the movie will also be more popular in general among audiences. In the linear model, it shows that there is close to 81% of variance in movie rating that can be explained by the popularity of movies. Popularity is statistically significant predictor for the vote_average.

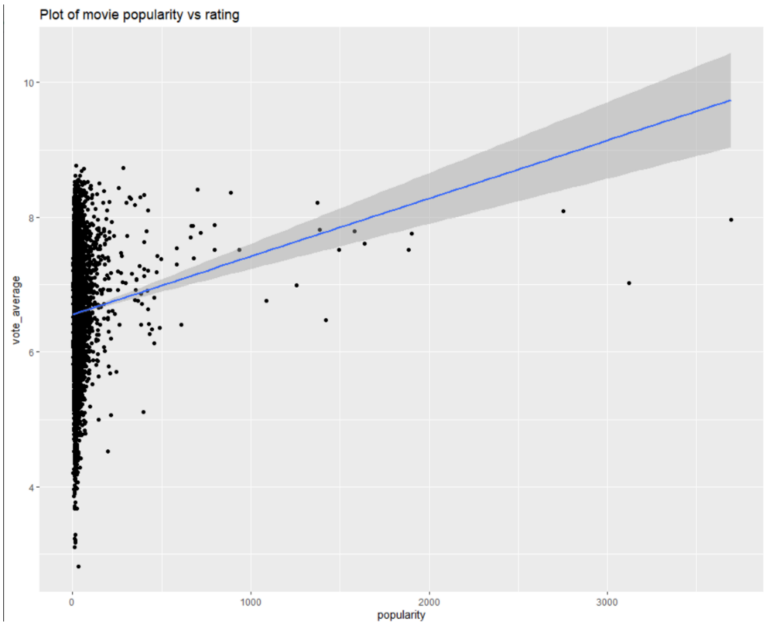


Figure 15: Regression Plot of Movie Popularity vs Vote_Average

The plot of movie budget and rating visualised the negative linear relationship between two variables.

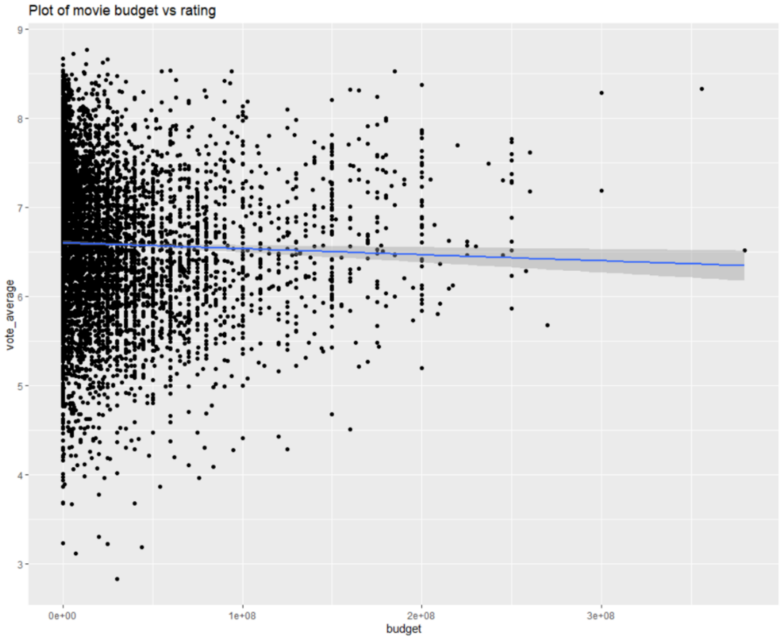


Figure 16: Regression Plot of Movie Budget vs Vote_Average

While Figure 17 shows a positive linear relationship between movie rating and revenue. The higher the movie rating is, the higher its box office earnings.

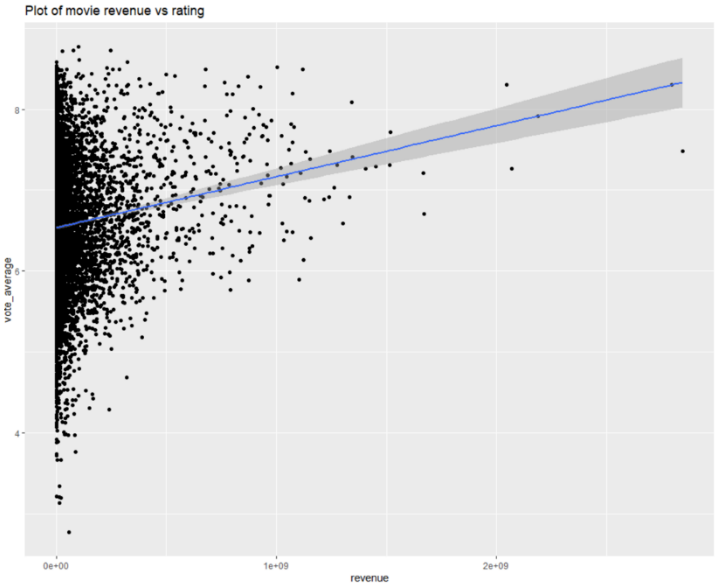


Figure 17: Regression Plot of Movie Revenue vs Vote_Average

There is a significant positive linear relationship between movie runtime and rating. The longer the movie, the higher its rating.

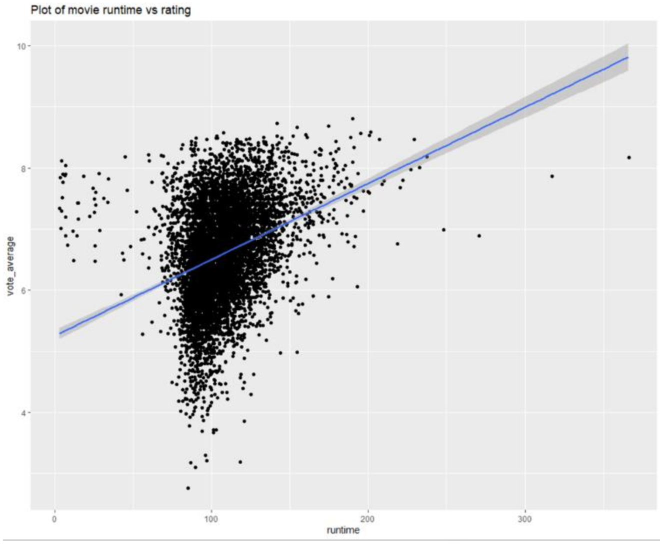
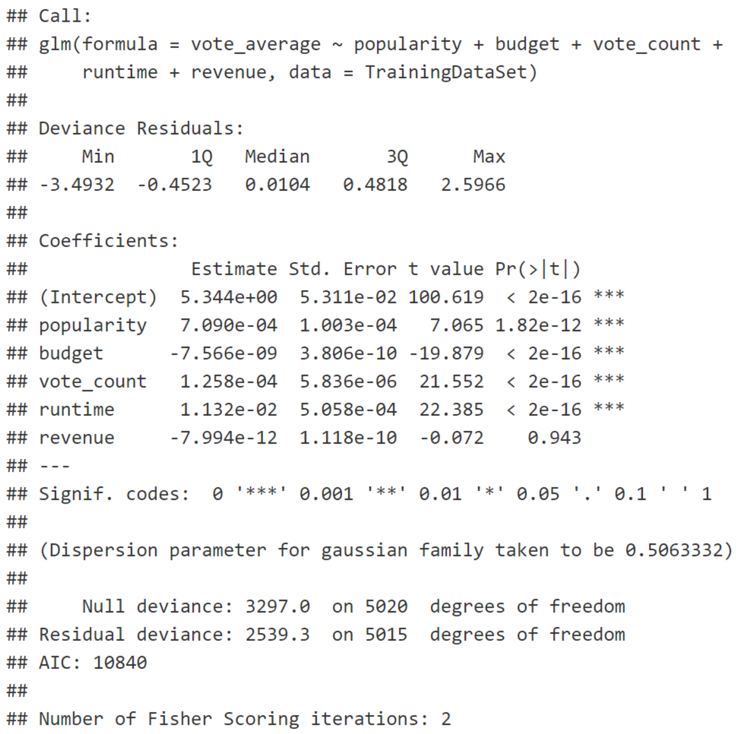


Figure 18: Regression Plot of Movie Runtime vs Vote_Average

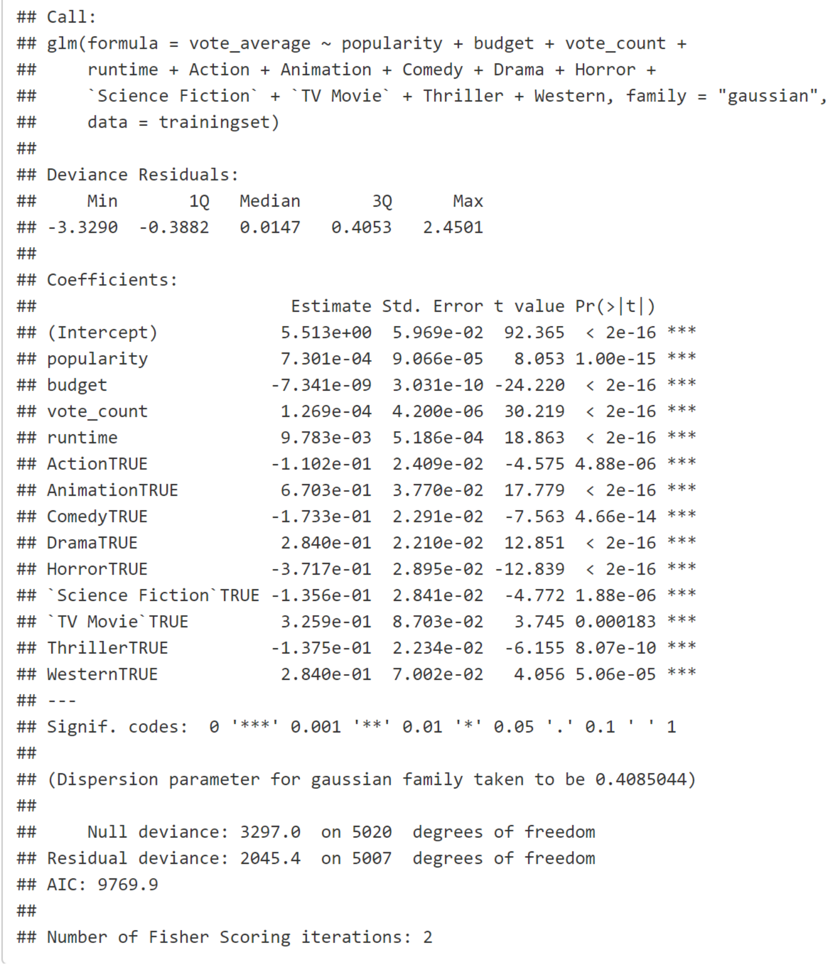
Generalized Linear Model (GLM)

For further model evaluation analysis, we utilise GLMs. For this analysis, we have used the same test/train split as per the linear regression model above (70% train, 30% test). We generated two models, regarding vote_average as output variable, GLM 1 is with the 5 significant input variables, while GLM 2 includes various genre types of movies as input variables. In comparison, GLM 2 has a better Akaike's Information Criteria (AIC), since it has a lower AIC, it needs less information to predict with almost the exact same level of precision.



```
## Call:
## glm(formula = vote_average ~ popularity + budget + vote_count +
##     runtime + revenue, data = TrainingDataSet)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4932  -0.4523   0.0104   0.4818   2.5966
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.344e+00  5.311e-02 100.619  < 2e-16 ***
## popularity   7.090e-04  1.003e-04   7.065 1.82e-12 ***
## budget      -7.566e-09  3.806e-10 -19.879  < 2e-16 ***
## vote_count   1.258e-04  5.836e-06  21.552  < 2e-16 ***
## runtime      1.132e-02  5.058e-04  22.385  < 2e-16 ***
## revenue     -7.994e-12  1.118e-10  -0.072    0.943
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.5063332)
##
##     Null deviance: 3297.0  on 5020  degrees of freedom
## Residual deviance: 2539.3  on 5015  degrees of freedom
## AIC: 10840
##
## Number of Fisher Scoring iterations: 2
```

Figure 19: GLM 1 with 5 significant Input Variables



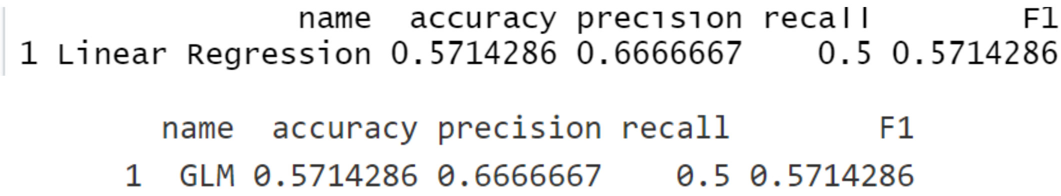
```
## Call:
## glm(formula = vote_average ~ popularity + budget + vote_count +
##     runtime + Action + Animation + Comedy + Drama + Horror +
##     `Science Fiction` + `TV Movie` + Thriller + Western, family = "gaussian",
##     data = trainingset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3290  -0.3882   0.0147   0.4053   2.4501
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.513e+00  5.969e-02  92.365  < 2e-16 ***
## popularity             7.301e-04  9.066e-05   8.053 1.00e-15 ***
## budget                -7.341e-09  3.031e-10 -24.220  < 2e-16 ***
## vote_count             1.269e-04  4.200e-06  30.219  < 2e-16 ***
## runtime                9.783e-03  5.186e-04  18.863  < 2e-16 ***
## ActionTRUE            -1.102e-01  2.409e-02  -4.575 4.88e-06 ***
## AnimationTRUE          6.703e-01  3.770e-02  17.779  < 2e-16 ***
## ComedyTRUE            -1.733e-01  2.291e-02  -7.563 4.66e-14 ***
## DramaTRUE              2.840e-01  2.210e-02  12.851  < 2e-16 ***
## HorrorTRUE            -3.717e-01  2.895e-02 -12.839  < 2e-16 ***
## `Science Fiction`TRUE -1.356e-01  2.841e-02  -4.772 1.88e-06 ***
## `TV Movie`TRUE         3.259e-01  8.703e-02   3.745 0.000183 ***
## ThrillerTRUE          -1.375e-01  2.234e-02  -6.155 8.07e-10 ***
## WesternTRUE            2.840e-01  7.002e-02   4.056 5.06e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.4085044)
##
##     Null deviance: 3297.0  on 5020  degrees of freedom
## Residual deviance: 2045.4  on 5007  degrees of freedom
## AIC: 9769.9
##
## Number of Fisher Scoring iterations: 2
```

Figure 20: GLM 2 with Genre Type Input Variables

All the input variables positively related with vote_average (movie rating), except movie budget and several movie genres would have negative linear relationship with movie rating as above linear regression model findings shown. Since the p value of the input variables are less than 0.05, THIS means the association between input and output variable are statistically significant.

Evaluation

Here, we are using confusion matrix, precision, recall and F1 as the evaluation measures to assist choosing the best model. As below figures show, two models indicated the same evaluation measure results, having 57% of proportion of correct predictions (accuracy), the correct proportion of target class predictions (precision) is 67%. While the proportion of the actual target class predicted correctly (recall) is 50%, and the blended metric of precision and recall score (F1) is around 57%.



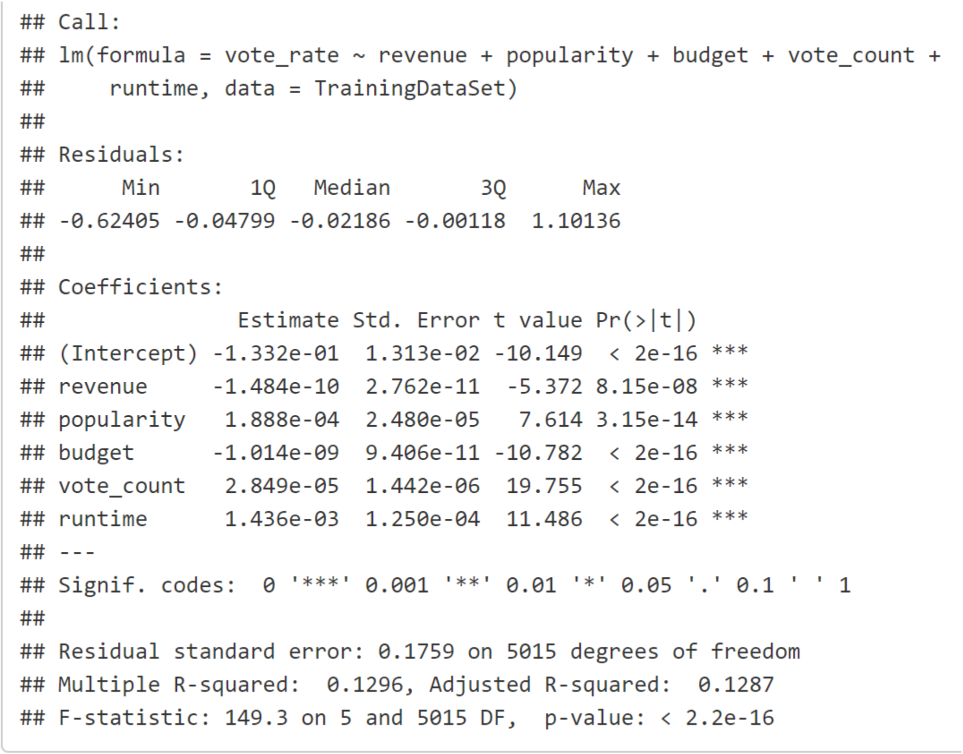
```
                 name  accuracy precision recall        F1
1 Linear Regression 0.5714286 0.6666667    0.5 0.5714286
```

```
  name  accuracy precision recall        F1
1  GLM 0.5714286 0.6666667    0.5 0.5714286
```

Figure 21: Evaluation Measure of Linear Regression & GLM

To obtain a better accuracy of the model, we create a new variable “vote_rate” and using a backward stepwise model selection method to limit the input variables to find the “best model fit”. The new variable “vote_rate” is found by taking the vote_average and assigning a binary of 0 and 1 where vote_average >= 8 is given a 1, and everything else is given a 0.

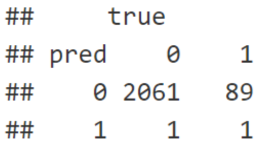
After adding new variable, as Figure 22 depicted that in the new linear regression model, the R-squared figure became lower from 71% to 18%, means the input variables can only explained 18% variability of output variable (vote_rate).



```
## Call:
## lm(formula = vote_rate ~ revenue + popularity + budget + vote_count +
##     runtime, data = TrainingDataSet)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62405 -0.04799 -0.02186 -0.00118  1.10136
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.332e-01  1.313e-02 -10.149  < 2e-16 ***
## revenue     -1.484e-10  2.762e-11  -5.372 8.15e-08 ***
## popularity   1.888e-04  2.480e-05   7.614 3.15e-14 ***
## budget      -1.014e-09  9.406e-11 -10.782  < 2e-16 ***
## vote_count   2.849e-05  1.442e-06  19.755  < 2e-16 ***
## runtime      1.436e-03  1.250e-04  11.486  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1759 on 5015 degrees of freedom
## Multiple R-squared:  0.1296, Adjusted R-squared:  0.1287
## F-statistic: 149.3 on 5 and 5015 DF,  p-value: < 2.2e-16
```

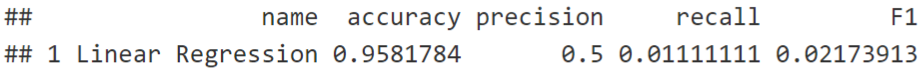
Figure 22: New Linear Regression Model with Vote_Rate

From the confusion matrix, there are none of the aforementioned "false negatives", but the linear regression model incorrectly predicts 82 of the 83 positives. The proportion of accuracy is higher from 57% to approximately 96% after adding new variable. However, the correct proportion of target class predictions dropped from 67% to 50%. And the proportion of the actual target class predicted correctly is lower from 50% to 1.2%. And the new model has lower F1, dropping to around 2.2%.



```
##     true
## pred    0    1
##    0 2061   89
##    1    1    1
```

Figure 23: Confusion Matrix of the New Linear Regression Model

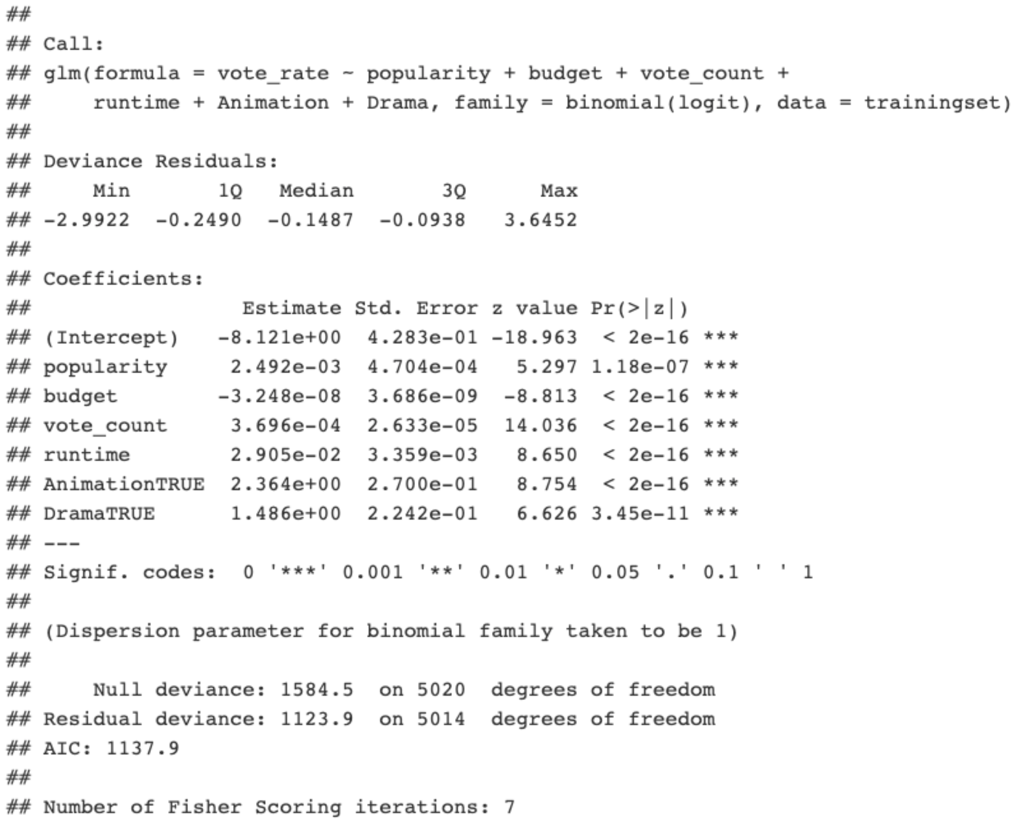


```
##                name  accuracy precision     recall         F1
## 1 Linear Regression 0.9581784       0.5 0.01111111 0.02173913
```

Figure 24: Evaluation Measure of the New Linear Regression Model

After using the backward stepwise selection method and refining the GLM model with selected variables using the new variable, vote_rate, as the output variable, we have been able to increase the accuracy and precision of our model to 96% and 72% respectively (as seen in Figure 27). This shows that there is a high precision in predicting which movie is classified as a high-performing movie. The AIC of the new GLM is also lower than GLM2's AIC of 9769.9. Having a relatively low AIC shows that the new GLM is a better model fit than that of GLM1

(10840) and GLM2. It should also be noted that we added the binomial(logit) family to this GLM model to account for the added vote_rate variable.

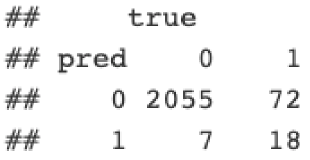


```
##
## Call:
## glm(formula = vote_rate ~ popularity + budget + vote_count +
##     runtime + Animation + Drama, family = binomial(logit), data = trainingset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9922  -0.2490  -0.1487  -0.0938   3.6452
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -8.121e+00  4.283e-01 -18.963  < 2e-16 ***
## popularity     2.492e-03  4.704e-04   5.297 1.18e-07 ***
## budget        -3.248e-08  3.686e-09  -8.813  < 2e-16 ***
## vote_count     3.696e-04  2.633e-05  14.036  < 2e-16 ***
## runtime        2.905e-02  3.359e-03   8.650  < 2e-16 ***
## AnimationTRUE  2.364e+00  2.700e-01   8.754  < 2e-16 ***
## DramaTRUE      1.486e+00  2.242e-01   6.626 3.45e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1584.5  on 5020  degrees of freedom
## Residual deviance: 1123.9  on 5014  degrees of freedom
## AIC: 1137.9
##
## Number of Fisher Scoring iterations: 7
```

Figure 25: New Generalized Linear Regression Model

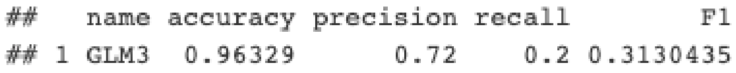
As can be seen from Figure 25, the following coefficients were of high significance to the vote_rate:

- Popularity, vote_count, runtime, Animation, and Drama, all of which has a positive influence on vote_rate.
- Budget, which has a negative influence on vote_rate, depicts that a larger budget does not automatically mean a movie will be of a highly rated.



```
##      true
## pred    0    1
##     0 2055   72
##     1    7   18
```

Figure 26: Confusion Matrix of the New Generalized Linear Regression Model



```
##   name accuracy precision recall        F1
## 1 GLM3  0.96329      0.72    0.2 0.3130435
```

Figure 27: Evaluation Measure of the New Generalized Linear Regression Model

The new GLM regression model has the F1 score of 31%, as can be seen in Figure 27. This could potentially be due to the imbalanced dataset. It would be recommended, should you choose to engage our services, for the dataset to be balanced with the use of the SMOTE() function as another step prior to further regression modelling to occur.

Findings & Recommendations

Film Production

Given the lack of positive correlation between film budget and vote average observed, we can advise that studios optimising for critical success can do so without larger budgets. This complements well with the positive linear relationship between budget and revenue, meaning that even films with a low budget will earn a similar profit relative to higher budget films. As such, smaller productions can be used to achieve high ratings while still seeing a comparable return on investment.

Conversely, studios are not recommended to rely on budget size to increase the likelihood of strong movie ratings. While films with large budgets were shown to experience proportionate amounts of revenue, the investment had no bearing on ratings. Ratings in particular affect longtail performance with home video purchases and online streaming platforms and as such, budget should not be employed as a success factor.

Stronger positive influencers for success were the genres of drama and animation, as well as longer runtimes. Films with these attributes were observed with higher ratings and higher popularity on TMDB.

Movies that are lower budget, with longer runtimes and are dramas or animated should be given particular consideration when greenlighting.

Film Marketing

Popularity online, particularly on TMDB, appears to create a virtuous circle of sorts, with greater popular, critical and commercial success all being associated with one another.

We recommend film studios capitalise on their strengths and market films that are internally considered to be strong performers more heavily, rather than bolstering weak performers.

We recommend that marketing teams utilise platforms such as TMDB and IMDB in order to spur greater revenues and critical performance.

Film Streaming

To surface more relevant films and increase viewing time, we recommend that streaming platforms utilise popularity data and rating data when user-generated heuristics are incomplete. That is, our research could be used for new films or new users where preferences have not been learned via user behaviour. Given the shared demographic of online users, the choice of TMDB and IMDN ratings is a relatively reliable proxy for preferences on these platforms.

Further Opportunities

Our team began with further research questions that we did not have time to fully answer. We recommend further analysis on them. They are below.

- Is there a relationship between a starring actor's popularity and movie ratings?
- Is there still a relationship between actors, genres, runtimes and movie performance when controlling for a budget?
- Does the death of a starring actor improve a movie's ratings?

Additionally, if these companies engage our services, we intend expand on our linear regression findings using logistic regression models to more thoroughly predict trends rather than describing and explaining them.

We also anticipate the use of classification models to result in more meaningful insights. Considering the contrast between foreign animated indie films and Hollywood blockbusters as one example, it would be particularly conducive to analyse k-means clusters and their separate trends.

Conclusion

As our analysis indicates, investigating a large dataset provides many opportunities to improve a film's critical and commercial success as well as its popularity.

This report is limited in scope and we recommend further exploration in terms of subject matter, different datasets, further research questions and other quantitative methods. For example, using a neural network to gain insights from the dataset may show more insights. Further, we recommend that this quantitative analysis not be used on its own or decontextualised but combined with qualitative analysis of films, preferably early in their production.

For more information, please see the data code samples linked in our references or reach out to the team.

Appendix

1. References

Amatriain, X., 'Mining large streams of user data for personalized recommendations', *Association for Computing Machinery Special Interest Group on Knowledge Discovery in Data Newsletter* 14:2, (2012), pp37-48.

Amazon, Amazon Studios, (2021), Website, https://studios.amazon.com (accessed 26 September 2021).

Au, R. Movie Madness, (2021), Github Repository, https://github.com/ronvoluted/moviemadness (accessed 3 September 2021).

Breidbach, CF., Maglio, P 'Accountable algorithms? The ethical implications of data-driven business models' *Journal of Service Management* (2020), pp163-185.

Gupta, N. et al. 'Capturing the Stars: Predicting Ratings for Service and Product Reviews.' *HLT-NAACL* (2010).

Markham, AN., Tiidenberg, K., Herman, A., 'Ethics as Methods: Doing Ethics in the Era of Big Data Research—Introduction', *Social Media and Society* (2019), pp1-9.

Netflix, Only on Netflix | Netflix Official Site, (2021), Website, https://www.netflix.com/au/browse/genre/839338 (accessed 26 September 2021).

Sparviero, S., 'Hollywood Creative Accounting: The Success Rate of Major Motion Pictures' *Media Industries Journal* 2(1), (2015), pp19-36

The Internet Movie Database, (2021), API https://developer.imdb.com/ (accessed 1 September 2021).

The Internet Movie Database, (2021), Datasets https://www.imdb.com/interfaces/ (accessed 1 September 2021).

The Movie Database, (2021), API https://developers.themoviedb.org/3/getting-started/introduction (accessed 1 September 2021).

Tomaselli, V., Giulio GC., and Valeria M.. ‘Complex Features in Review Bombing’, *Book of Abstracts* (2020) p. 18. 2020.

Tripathy, JP., ‘Secondary data analysis: Ethical issues and challenges.’ *Iranian journal of Public Health* 42(12) (2013), pp1478-1479.

2. Data Code Samples

Datasets used are completely reproducible and our team’s GitHub Repository is available here (Au 2021) where code can be viewed in full.

Data acquisition processes

Function to query TMDB API top-rated-movies endpoint (get-top-rated-movies.R – line 26)

```
get_top_rated <- function(page = 1, apikey = keyring::key_get('tmdb_api
_key')) {
  response <- httr::GET("https://api.themoviedb.org/3/movie/top_rated",
                        query = list(api_key=apikey, page=page))

  httr::content(response, 'parsed')[["results"]]
}
```

Function to build and save datasets (get-top-rated-movies.R – line 35)

```
save_top_rated <- function(range=5, start=1, filename="top_rated_movies
.tsv") {
  ...

  # Create empty TSV file with only column headers
  write.table(columns_appended, filename, sep="\t", row.names=FALSE, co
l.names=FALSE, quote=FALSE)

  # Request a page of results for the entire range
  for (i in start:(start + range - 1)) {
    results_page <- get_top_rated(i)

    for (result in results_page) {
      ...

      # Append result to TSV. Tab separated values due to heavy use of
commas in dataset
      write_delim(tibble(result), filename, delim = "\t", append = TRUE)
    }
    # While TMDB explicitly places no restrictions, rate limit the requ
est out of courtesy
    Sys.sleep(5)
  }
}
```

Building full dataset from all TMDB entries (get-top-rated-movies.R – line 129)

```
# Save every top rated movie from TMDB to a TSV file. Make some tea! --
--

check_top_rated()[["total_pages"]] %>%
  save_top_rated(filename="data/all_top_rated_movies.tsv")
```

Data merger processes

Function to query TMDB API movie-details endpoint (get-movie-details.R)

```
get_movie_details <- function(id, apikey=keyring::key_get('tmdb_api_key
')) {
  response <- httr::GET("https://api.themoviedb.org",
                        path=list("3", "movie", id),
                        query=list(api_key=apikey))

  httr::warn_for_status(response, task="GET movie details")

  httr::content(response, 'parsed')
}
```

Merge movie-detail headers into top-rated-movies headers (get-top-rated-movies.R – line 44)

```
columns <- list("id", "title", "original_title", "overview", ...)

genres <- list("Action", "Adventure", "Animation", ...)
columns_appended <- c(columns, "tagline", "runtime", "budget", "revenue
", genres)
```

Merge movie-detail data into top-rated-movies data (get-top-rated-movies.R – line 69)

```
# Get further movie details from separate endpoint
details <- get_movie_details(result[["id"]])

# Append expanded details to current result

result["tagline"] <- details[["tagline"]]
result["runtime"] <- details[["runtime"]]
```

```r
result["budget"] <- details[["budget"]]
result["revenue"] <- details[["revenue"]]

genre_names <- lapply(details[["genres"]], function(genre) {
  genre[["name"]]
})
```