Final Report: IMDb Movies Dataset

Adam Rychtecky Erin Smith Ali Shahid Miao Qi
Matthew Arteaga

2025-03-05

Table of contents

1. Abstract

This Report investigates how multiple linear regression can be applied to analyze the relationship between a movie's Grossing World Wide with five key predictor variables (Question 1), as well as the relationship between the number of Drama and Action movies released with the year that they were released (Question 2). For each of these questions, data was cleaned to remove potential outliers and or observations that lacked necessary information. From there, the data was initially explored by analyzing summary statistics of the variables of interest as well as preliminary plots to examine the linearity of the data. For Question 1, Stepwise Variable selection was utilized as well as a Box-Cox transformation. For each question a Linear Model was produced and it's overall significance as well as individual significance for each predictor variable was analyzed. The models produced for both questions were found to be statistically significant, supporting the initial hypotheses proposed.

2. Introduction

2.1 Dataset Overview

The dataset, sourced from Kaggle, consists of movies released between 1960 and 2024. Key variables include:Budget (production cost), Gross Worldwide Revenue (total earnings), Rating (IMDB audience score), Votes (number of votes), Nominations (number of award recognition), Duration (runtime in minutes), Year of Release (temporal aspect), genres (genres in movie), countries_origin (countries in filming).By examining these social and economic factors of the movie making industry we aim to make significant inferences of movie success and attributes in the industry.

2.2 Research Questions and Hypotheses

The film industry has undergone significant transformations over the years, with advancements in technology, changing audience preferences, and evolving economic factors shaping movie production and success. This study explores two key questions related to movie success and industry trends using multiple regression analysis.The first being, what are the best indicators for a blockbuster movies success in the form of gross world wide income? We hypothesize that budget will have the highest effect on income due to it contributing to the creation of economically successful movies. Secondly, we wonder how has movie duration evolved over time and can a movie's genre help predict average movie duration? We hypothesize that movie duration has slowly been increasing over time as technology advances, and that the rate of increase differs depending on the genre.

2.3 Data Preparation

2.3.1 Data Cleaning

Data cleaning and preprocessing were critical steps in preparing the dataset for regression analysis. The original dataset of every movie between the years of 1960 to 2024 was extremely large with lots of missing data. We reasoned that most of the missing data was from small unkown movies that did not fit the agenda of the research study. We also found that budget was not standardized to USD. To get around this we switched the aim of the research study to only United States made movies. This greatly reduced the number of observations and missing data. Removing the rest of the missing data left us with just over 7000 observations. Then for each research question we further cleaned the data to fit the objective of the research.

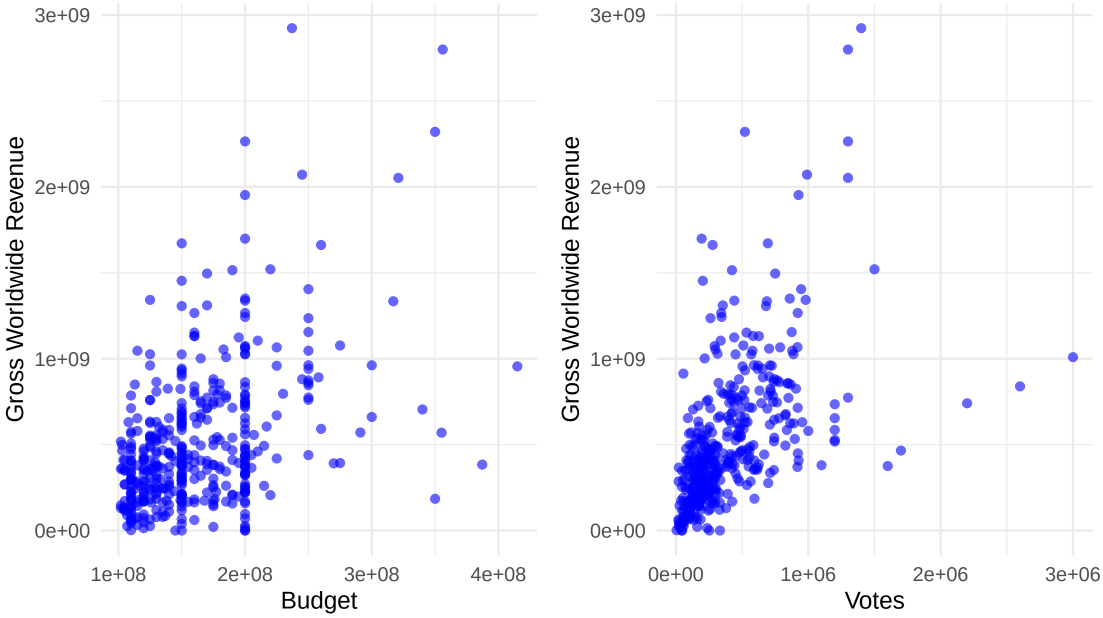
(Identifying outliers)

2.3.2 Exploratory Data Analysis (EDA)

Summary Statistics of Key Variables:

```
     Rating          budget              Votes          total_minutes
 Min.   :3.80   Min.   :1.02e+08   Min.   :   4300   Min.   : 76
 1st Qu.:6.10   1st Qu.:1.25e+08   1st Qu.: 165250   1st Qu.:106
 Median :6.60   Median :1.50e+08   Median : 281000   Median :123
 Mean   :6.65   Mean   :1.62e+08   Mean   : 392923   Mean   :124
 3rd Qu.:7.30   3rd Qu.:1.94e+08   3rd Qu.: 522500   3rd Qu.:138
 Max.   :9.00   Max.   :4.15e+08   Max.   :3000000   Max.   :206
  nominations    grossWorldWide
 Min.   :  0.0   Min.   :3.37e+04
 1st Qu.: 10.0   1st Qu.:2.38e+08
 Median : 22.0   Median :4.07e+08
 Mean   : 37.9   Mean   :5.14e+08
 3rd Qu.: 48.0   3rd Qu.:6.81e+08
 Max.   :425.0   Max.   :2.92e+09
```

Visualizations:



Both the scatterplots of `grossWorldWide` vs. `budget` and `grossWorldWide` vs. `Votes` indicate a likely linear positive relationship between the respective predictor variables and our selected response variable.

3. Regression Modeling Process

3.1 Model Selection

As part of our analysis, we first conducted an exploratory data analysis (EDA) to assess the distribution of key variables. A boxplot analysis of the budget variable revealed it had the highest spread and the greatest number of outliers. Given the large variance in production budgets across films, we chose to focus our analysis on blockbuster-level movies, defined as those with a high production budget. This subset refinement reduced the dataset to 422 observations, allowing us to analyze films with a comparable scale of investment and revenue potential. After initial tests of different candidate models 4 stood out among all trials given below.

In order to investigate whether multicollinearity exists between the five selected predictor variables, we can analyze the VIF values produced from the initial model including the six key variables: grossWorldWide, Rating, budget, Votes, total_minutes, and nominations in which grossWorldWide is the response variable (in USD) and the rest are predictor variables.

Based on the VIF values, because none of them even exceed a value of three, it is fair to assume that there is no concern of multicollinearity between the predictor variables. This is consistent with the investigation of the correlation matrix.

To determine the most appropriate predictive model for grossWorldWide, we applied the Box-Cox transformation to the response variable across four candidate models, each yielding its respective optimal lambda value. Following the transformation, we evaluated the models based on key selection criteria, including AIC, BIC, and Adjusted R². Model 2, which included Votes, budget, and Rating as predictors, demonstrated superior performance in terms of model fit and explanatory power, making it the preferred choice.

Name	AIC	BIC	R2_adjusted	RMSE
modelA_transformed	7520.458	7536.638	0.4123683	1775.928
modelB_transformed	7510.195	7530.420	0.4278340	1750.310
modelC_transformed	7512.042	7536.312	0.4266689	1749.994
modelD_transformed	7511.760	7540.075	0.4283910	1745.267

To ensure the robustness of our results, we conducted an outlier analysis, identifying approximately 6% of the observations as influential points based on Cook's Distance and leverage diagnostics. After removing these observations, we re-estimated the model, leading to substantial improvements in overall model performance.

3.2 Model Results

After performing A Stepwise selection on the transformed data, the key variables selected for the model include: `budget`, `Votes`, and `Rating`. After removal of influential points based on Cook's distance and leverage diagnostics, the best model for predicting `grossWorldWide` comes out to be:

$$g\hat{W}W = 1.543 * 10^3 + 1.574 * 10^{-5} b + 3.833 * 10^{-3} V + 2.481 * 10^2 R$$

Where:

- $g\hat{W}W$: grossWorldWide in USD
- b: budget in USD

- V: number of votes on IMDb
- R: IMDb Rating

Table 1: Coefficients and Significance

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1542.6648722170	847.569674909532	1.820104	0.06951951567437775203
budget	0.0000157349	0.000002036491	7.726473	0.0000000000009695960
Votes	0.0038331693	0.000409270962	9.365847	0.0000000000000000065
Rating	248.0812149007	134.066298219206	1.850437	0.06501651847875145052

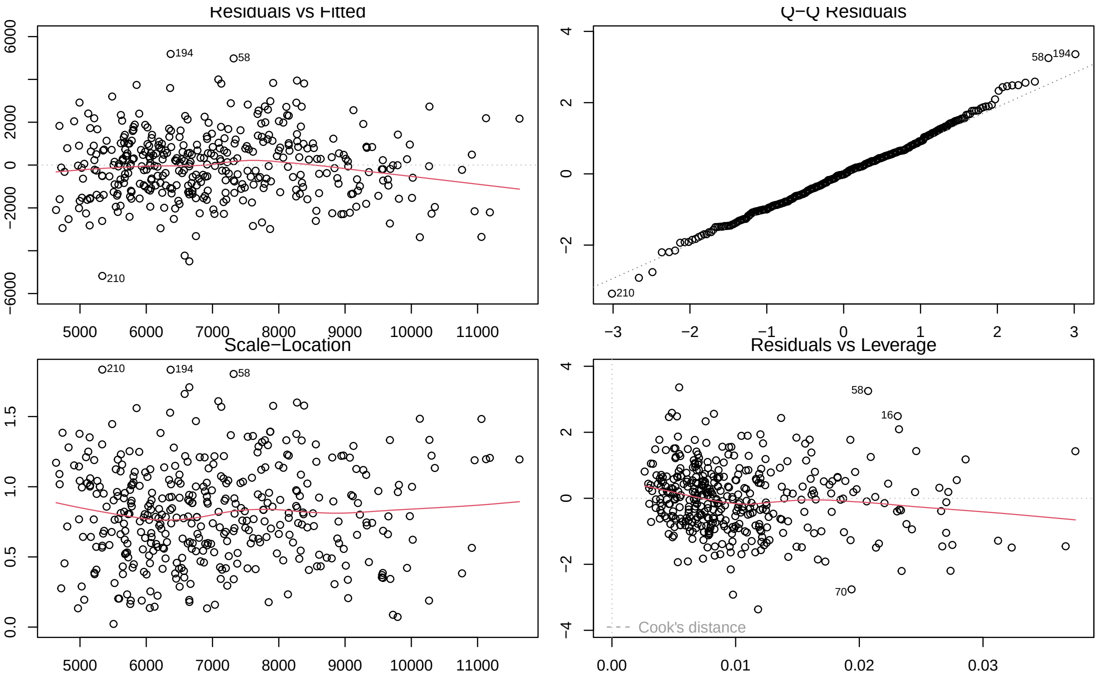
The p-values generated from the linear model indicate that at $\alpha = 0.05$, we reject the null hypothesis that the coefficients for the predictor variables: `budget` and `Votes` are equal to zero, thus they have a meaningful impact on `grossWorldWide` in this model. However at $\alpha = 0.05$, we fail to reject the null hypothesis that the coefficient for the `Rating` predictor variable is equal to zero.

Interpretation of Significant Predictors:

- `budget` coefficient: for every one dollar increase to `budget`, `grossWorldWide` is expected to increase by $1.574 * 10^{-5}$ dollars.
- `Votes` coefficient: for every additional vote on IMDb, `grossWorldWide` is expected to increase by $2.481 * 10^{2}$ dollars.

3.2.1 Assumptions

A subsequent assessment of regression assumptions was conducted to validate the reliability of the refined model in predicting grossWorldWide. The Shapiro-Wilk test for normality yielded a p-value of 0.098, confirming that the residuals follow a normal distribution (assumption met). The Breusch-Pagan test for homoscedasticity returned a p-value of 0.167, indicating constant variance of residuals (assumption met). Additionally, the Durbin-Watson test for independence of errors produced a p-value of 0.228, confirming that autocorrelation was not present in the residuals (assumption met). In addition to these numeric tests plots for each are shown below. These results collectively validate the suitability of the model, ensuring compliance with the fundamental assumptions of linear regression.



3.3 Recommendations

Based on the findings of our model, specifically when investigating the significance of our coefficients, one improvement that could be made to our model would be to remove the `Rating` predictor variable as it is not statistically significant at $\alpha = 0.05$. Additionally, to improve our model, we might want to investigate other predictor variables outside the scope of our dataset to improve our adjusted R^2 value, which is `0.466` with our current model.

Conclusion

Our Initial hypothesis was that a movie's Grossing World Wide could be predicted with a linear combination of the selected key variables: `budget`, `Rating`, `Votes`, `total_minutes`, and `nominations`. Based on the regression model, this hypothesis was partially supported as we found that only `budget`, `Rating`, and `Votes` were required for the model of best fit based on Stepwise variable selection. Upon Analysis of the p-values yielded from the Linear Model, we found that the `Rating` predictor was statistically insignificant in the model. With the initial investigation of the VIF values produced from the full model, no concern of multicollinearity was found so no Ridge Regression was warranted for the model.

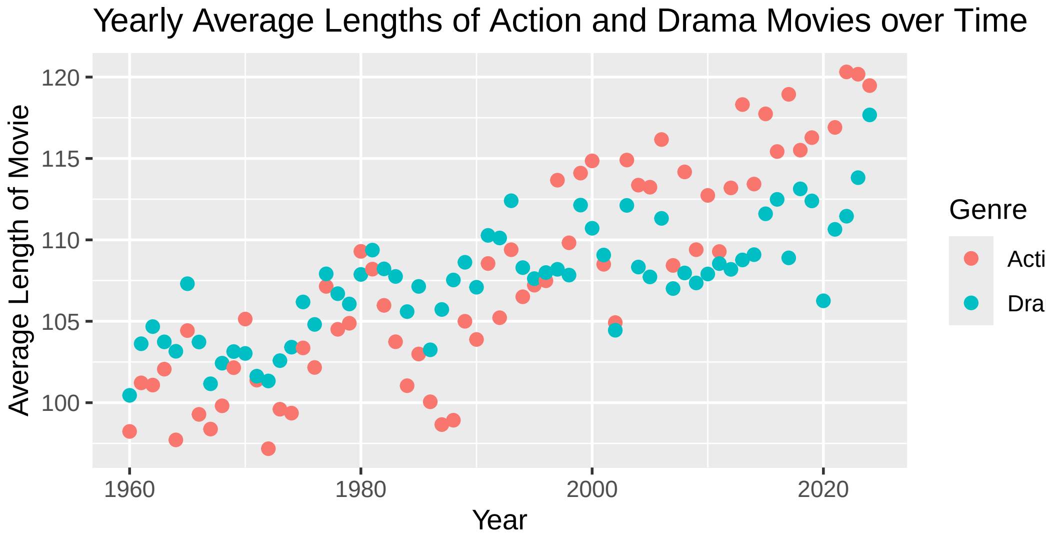
4. Hypothesis 2

4.1 Data Preparation

We used the full IMDB dataset spanning from 1960 to 2024 and cleaned it so that it only contained the relevant variables: Year, Duration (minutes), and Genre. For genre, only the relevant genres (drama or action) were kept. To account for movies that fell into both of these categories we duplicated them and put one entry in Drama and the other in Action. These genres were chosen for their popularity, giving a large sample size even with the overlap between genres. On the cleaned dataset, we cleared any entries missing values or movies with duration's equal to zero. Several outliers were immediately noticed in the form of movies that were unusually long (24+ hrs). These were subsequently removed from the dataset. After cleaning individual outlier cases, the yearly average duration for each genre was taken .

4.2 Exploratory Data Analysis (EDA)

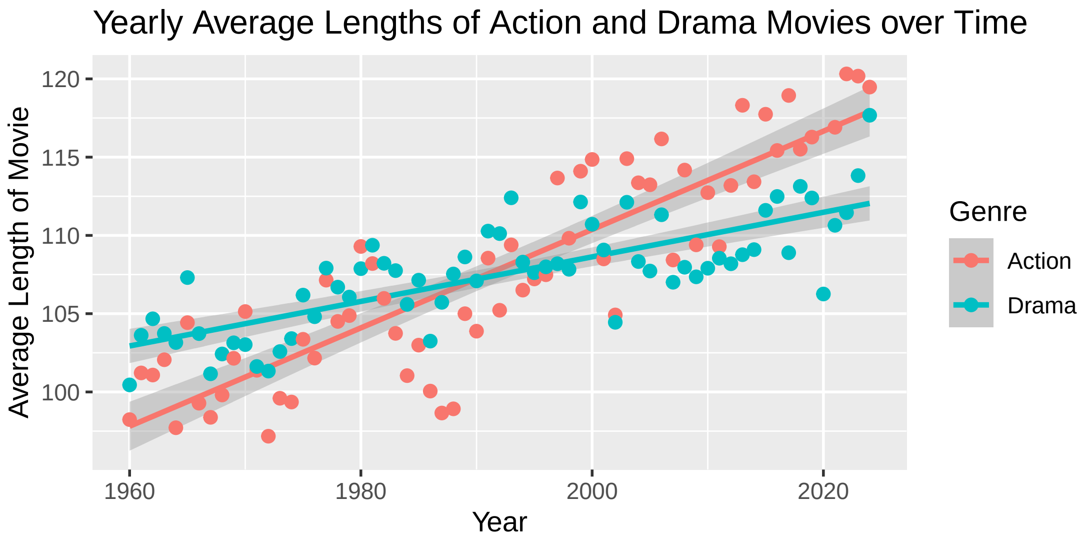
Initially, we were curious as to how the duration of each genre may have varied over time. To view the change over time in a simple format, we used a double scatterplot to show how the two genres' average lengths changed against each other. As can be seen below, action movies appear to have a sharp increase in average length over time, while drama movies had a much softer increase. From the summary statistics we learn that all predictors as well as their interactions are significant at a 0.001 significance level and the overall model has a p-value of 2.2e-16 and an F-statistic of 117.5 on 3 and 125 df. This suggests much of the error seen in the model is explained by the regression model. Furthermore, the adjusted r-squared is 0.73 which also supports the assertion that much of the variation is explained by the residuals.



4.3 Regression Model

$$Duration_{i\lambda} = -517.9224 + 0.3141 \times Year + 341.9127 \times GenreDrama$$
$$- 0.1718 \times Year : GenreDrama + \epsilon_i$$

To adjust for the effect of Year on Genre, we included the interaction between these predictors in our linear model, and because the interaction was significant at a 0.01 level (p = 2.64e-11 < 0.01). Hypothesis testing found all predictors (Year, Genre) to be significant in predicting the average duration of a genre (p < 0.01). Action is the reference genre, and based on the regression it appears that the genre Drama starts off with a higher average length, but sees a slower growth rate in duration than Action. This is reflected in a graph of our model.

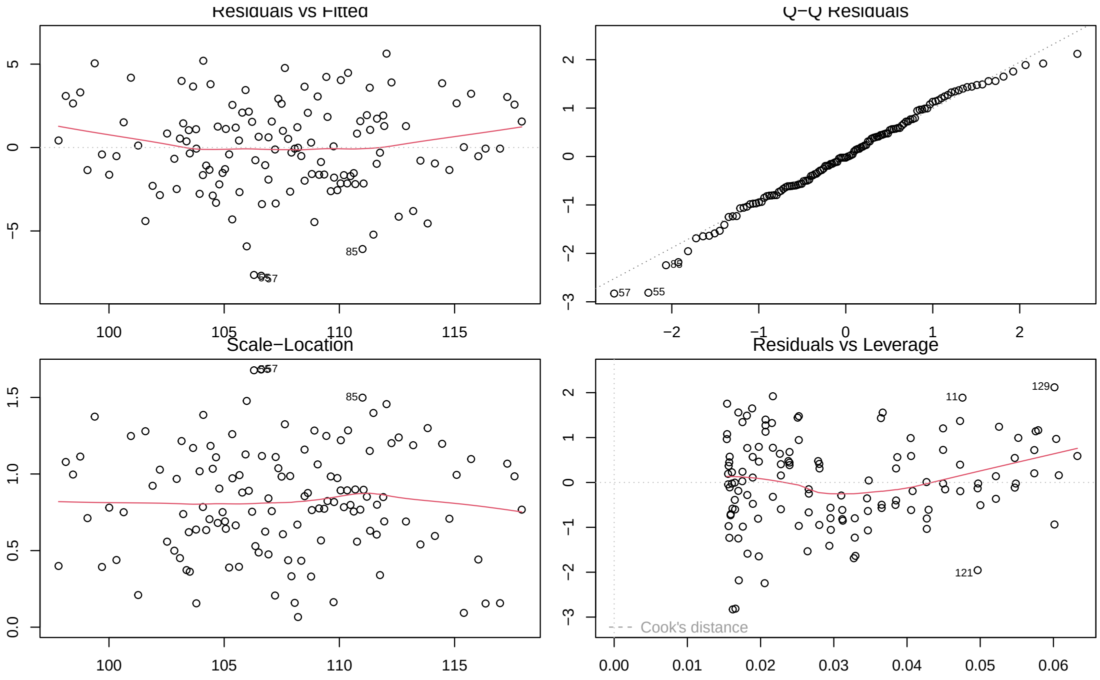


No transformation was applied in the end as they did not significantly improve the model, and just decreased the interpretability of the results.

The regression model has four coefficients:

1) $Intercept$ (-517.9224) is a negative and somewhat nonsensical value since there is no year zero.
2) $Year$ (0.3141) the yearly increase in movie duration for Action movies.
3) $GenreDrama$ (341.9127) is a dummy variable that adjusts the intercept if the genre of the movie is drama. It indicates that Drama movies started off as having a higher average duration than Action movies.
4) $Year : GenreDrama$ (-0.1718) adjusts for the interaction between Drama movies and year.

4.3.1 Assumptions



1) Heteroscedasticity: A plot of the residuals vs predicted revealed random scatter about the line, suggesting homoscedasticity of the residuals. A Breusch-Pagan test reveals there is not enough evidence at a 0.01 significance level (0.03468 > .01) to say the variance is non-constant (heteroscedastic).

2) Normality: The normality assumption for the linear regression is met and we fail to reject the null hypothesis for the Anderson-Darling normality test (p = 0.7512 > 0.05). There is evidence to suggest the residuals of the regression model do follow an approximate normal distribution. Examining the Q-Q plot, we see points roughly fall along the line which supports the normality assumption.

3) Independence: For autocorrelation, the p-value is not significant at a 0.001 significance level. The DW value is 1.4831, which is closer to two than zero or four. This suggests that while it is a positive autocorrelation, it is still relatively low and we still believe the model to be acceptable.

4) Linearity: Hypothesis testing revealed all predictors included in the model to be significant at a 0.05 significance level. This is supported by the scale-location plot, which has a relatively straight line and evenly distributed points.

4.4 Recommendations

The sample sizes of our two genres, despite both being popular shows a huge disparity in size. Action has a much smaller sample than Drama. This can cause unequal statistical power, and weakens the significance of the results. Furthermore, the scope of our study is limited. We only chose two genres for our analysis due to the high overlap between genres. Initially, we wanted to look at many genres, however, IMDB places movies in multiple categories such that it is extremely rare to find a movie that is in a singular category. We decided to only use two genres to minimize overlap while still providing large samples to draw from. Action and Drama were chosen for their popularity, and because they were two of the largest genres available.

Conclusion

Our hypothesis was that over time we would see an increase in movie length, and that we expected drama's to be longer on average than action movies. Based on the regression model, this hypothesis is supported. Looking further, our Regression model shows that in the 1990s action movies outpaced dramas in average length. We hypothesize that this may be due to the success of longer-form (2+ hours) action movies such as the Matrix or Jurrasic World franchises, leading to other studios trying to replicate this success with similar films. A closer analysis of the IMDB data and perhaps a cultural analysis could help to explain the shifts we see in the data.