C3 Group1
Group members: Yufan Liu, Marie Picini, Yushan Guo, Yutong Wu
Milestone 1

1. Project idea: What is the main goal of your project?
We want to explore the relationship between various variables within our dataset like genre, director, and budget to predict average movie score.
Dataset: https://www.kaggle.com/datasets/mohammedalsubaie/movies/data

2. Variables:
Dependent (response) variable: average scores

Independent (predictor) variables:
1) genre (qualitative)
2) popularity (quantitative)
3) budget (quantitative)
4) production companies (qualitative)
5) director (qualitative)
6) spoken_language (qualitative)
7) release_date (qualitative)
8) revenue (quantitative)
9) vote_count (quantitative)
10) Production_countries (qualitative)

3. Hypotheses:

We have many hypotheses and would like to explore many relationships and ways that various columns predict ratings within our dataset, but we have listed a few preliminary hypotheses below.

We hypothesize that higher revenue predicts higher movie ratings.
We hypothesize that the higher budget predicts higher movie ratings.
We hypothesize that certain genres, specifically action, will on average have higher movie ratings.
We hypothesize that greater popularity predicts higher movie ratings.

4. Contribution:

We will split all tasks evenly as they come up.

Milestone 2

1. Project idea: What is the main goal of your project?

We want to explore the relationship between various variables within our dataset like genre, director, and budget to predict average movie score. Although the dataset has a very low credibility as it did not publish any source/data collection information, it is still usable because the dataset is very complete with a 7.65 usability score on Kaggle.
Dataset: https://www.kaggle.com/datasets/mohammedalsubaie/movies/data

2. Variables:

By analyzing both qualitative factors like genre and production companies, as well as quantitative variables such as budget and revenue, we aim to uncover key predictors of a movie's success in terms of average rating scores.
Dependent (response) variable:

- vote_average (quantitative)
    - Average rating of the movie (from 0-10). This is our responsive variable, which would respond differently to the deterministic variables. e.g the higher revenues of a movie, the lower the average rating scores.

Independent (predictor) variables:

- genre (qualitative)
    - Different genres tend to appeal to varying audience preferences and expectations. Action movies may be rated highly by fans of thrilling, fast-paced plots.

- popularity (quantitative)
    - Popular movies tend to generate more buzz, resulting in higher viewership and reviews, which can influence ratings.

- budget (quantitative)
    - Higher budgets often allow for better production values and famous actors, so we assume higher budgets lead to higher rating scores.

- production companies (qualitative)
    - Some production companies are known for producing films that perform well at film festivals. We assume that relatively famous production companies would lead to higher rating scores.

- director (qualitative)
    - The director's visual style and the way the story is presented directly affects the audience's emotional resonance, which in turn affects the rating.
    - The director guides the actors' performances, and excellent performances can enhance the realism of the characters, which in turn affects the audience's evaluation of the movie.

- spoken_language (qualitative)
    - The use of language familiar to the audience can better convey the emotions and stories of the characters, making it easier for the audience to resonate and enhancing the viewing experience.
    - Different languages carry different cultural backgrounds, and movies using a specific language can make the audience feel a stronger sense of cultural identity and increase their enjoyment of the movie, thus improving the rating.

- release_date (qualitative)
    - Movies released during specific times of the year, such as holidays or summer, may receive more attention and higher viewership. Seasonal releases can attract larger audiences, influencing ratings.
- revenue (quantitative)
    - High revenue often indicates strong audience interest and appeal, which can lead to more positive reviews and ratings.

- Production_countries (qualitative)
    - This variable describes which countries the movie was originally produced in. Audiences may have preferences for a particular country's movie. e.g if audiences have a preference for American movies, then rating for American movies would be higher than other countries'.

3. Hypotheses:

We have many hypotheses and would like to explore many relationships and ways that various columns predict ratings within our dataset, but we have listed a few preliminary hypotheses below.

We hypothesize that higher revenue predicts higher movie ratings.
We hypothesize that the higher budget predicts higher movie ratings.

We hypothesize that certain genres, specifically action, will on average have higher movie ratings.
We hypothesize that greater popularity predicts higher movie ratings.

4. Sample Data:

Table:

Description: df [6 × 12]

	title <chr>	vote_average <dbl>	vote_count <int>	release_date <dbl>	revenue <dbl>	adult <lgl>	budget <int>	popularity <dbl>
1	Inception	8.364	34495	2010	825532764	FALSE	160000000	83.952
2	Interstellar	8.417	32571	2014	701729206	FALSE	165000000	140.241
3	The Dark Knight	8.512	30619	2008	1004558444	FALSE	185000000	130.643
4	Avatar	7.573	29815	2009	2923706026	FALSE	237000000	79.932
5	The Avengers	7.710	29166	2012	1518815515	FALSE	220000000	98.082
6	Deadpool	7.606	28894	2016	783100000	FALSE	58000000	72.735

6 rows | 1-9 of 12 columns

5.
95% Confidence Intervals:
vote_average: (5.993463 6.006672**)**
vote_count: (63.45697 67.51040)
revenue: (41652458 45523986)
budget: (10002848 10637007)
popularity: (2.636532 2.726043)

2021 average movie revenue in the U.S is $10,165,556
(https://www.boxofficemojo.com/year/)

Milestone 3

Understanding the factors that influence movie success has become increasingly important in today's era when the movie industry has become a major cultural and economic force. As the number of movies released each year continues to grow, filmmakers and companies are increasingly interested in predicting which movies will resonate with audiences. The purpose of this project is to explore the relationship between a variety of factors (e.g., genre, budget, and director) and a movie's average rating. We aim to analyze qualitative factors including genre and production company, as well as quantitative variables such as budget and revenue, in order to reveal key predictors of a film's success. Despite the fact that the dataset used is less credible in terms of source and collection methodology, we will utilize a comprehensive dataset from Kaggle that seeks to identify significant predictive variables that influence movie success.

Our analysis will focus on both qualitative and quantitative factors. We hypothesize that there may be a positive correlation between higher budgets and revenues and higher average ratings, as these factors typically imply higher production quality and stronger marketing. In addition, we will examine whether specific genres of movies, particularly action movies, receive higher ratings because they cater to audiences' need for thrill-seeking experiences. The role of directors will also be included in the study, as their unique narrative techniques may significantly affect audience engagement and emotional response. Also, we will explore how language and release date affect ratings, taking into account the significant impact of cultural identity and timing choices on audience ratings.

By articulating these hypotheses, we hope to contribute to an understanding of what factors make a movie successful in the eyes of the audience. Through this study, we expect to reveal the intricate relationships between these variables and provide insights that will help filmmakers make more informed decisions that will enhance the potential for movie success.

In our analysis, we will explore both qualitative and quantitative factors to uncover key predictors of a movie's success, as measured by average rating scores. The dependent variable in our study is the average rating of the movie, which ranges from 0 to 10, serving as a response variable that can be influenced by various independent (predictor) variables. Among these predictor variables, genre is a qualitative factor that reflects the diverse audience preferences and expectations; for example, action movies often receive higher ratings from fans who enjoy thrilling and fast-paced narratives. Popularity, measured quantitatively, is another important

factor, as popular movies typically generate more buzz, leading to increased viewership and reviews that can positively impact ratings.

Budget, also a quantitative variable, is assumed to correlate with higher average ratings due to the potential for better production values and the inclusion of well-known actors. The production company is a qualitative variable, with certain companies recognized for their successful films at festivals; we hypothesize that movies produced by more reputable companies will achieve higher ratings. Additionally, the director's influence is crucial, as their unique visual style and storytelling approach can significantly affect the audience's emotional response, thus impacting ratings. The spoken language of a film plays a qualitative role as well; movies in familiar languages may resonate better with audiences, enhancing their viewing experience and ratings. The release date, another qualitative factor, can influence ratings, as films released during peak seasons, like holidays or summer, often attract larger audiences. Furthermore, revenue serves as a quantitative indicator of a movie's appeal; high revenue typically reflects strong audience interest, which may translate into more positive reviews. Lastly, the production countries—indicating where the film was originally produced—may also affect audience preferences, potentially leading to higher ratings for films from countries that viewers favor, such as American movies.

In our analysis, we aim to explore various relationships between the predictor variables and movie ratings within our dataset. We propose several preliminary hypotheses that guide our investigation. First, we hypothesize that higher revenue is associated with higher movie ratings, as substantial financial returns often indicate strong audience interest and positive reception. Second, we believe that a higher budget will similarly predict elevated movie ratings, given that increased funding can enhance production quality and attract prominent actors. Additionally, we hypothesize that specific genres, particularly action films, will generally receive higher average ratings due to their appeal to audiences seeking thrilling experiences. Finally, we anticipate that greater popularity, measured by factors such as audience buzz and viewership, will also correlate with higher movie ratings. By examining these hypotheses, we hope to gain insights into the key factors that contribute to a film's success.

In our research, we will first clean the data by removing observations with a value of 0 and retaining only the independent variables and dependent variables we wish to explore. Next, we will use R to perform graphical analysis, including creating Q-Q plots and residuals versus variable plots, to visually assess the normality of the data and the linear relationships. We will interpret these plots for each variable to determine if any transformations are necessary and evaluate whether any variables should be dropped. Finally, we will identify outliers in the data and decide on appropriate actions to address them, ensuring a solid foundation for our analysis and supporting our paper's conclusions.

This is our sample data and the 95% confidence interval we get:

Description: df [6 × 12]

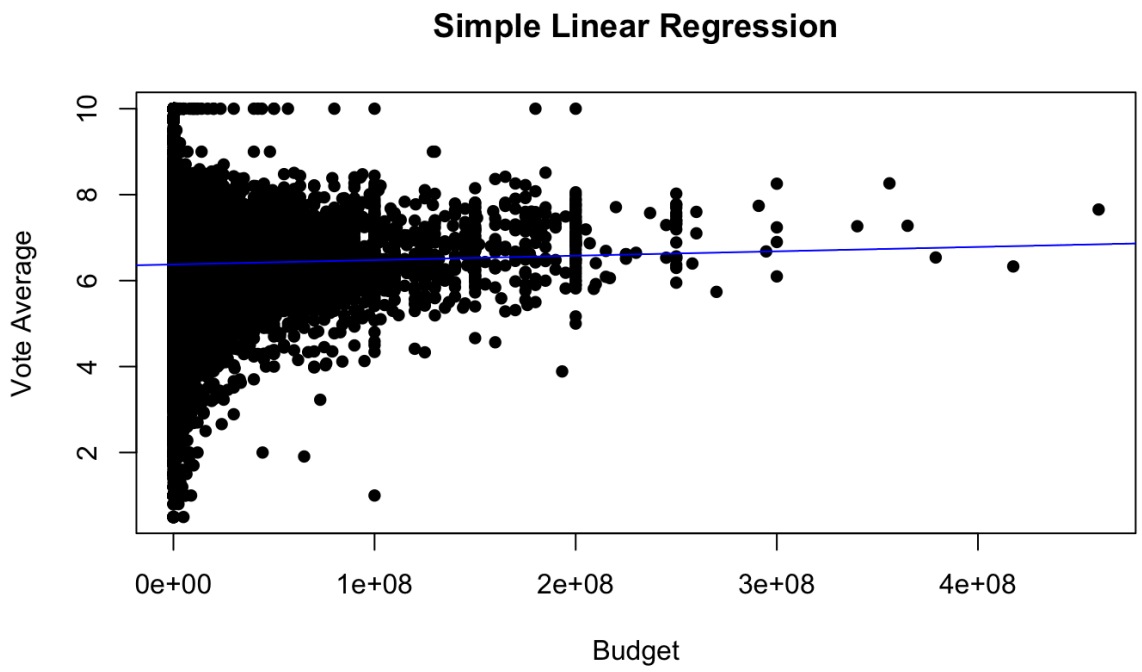
	title <chr>	vote_average <dbl>	vote_count <int>	release_date <dbl>	revenue <dbl>	adult <lgl>	budget <int>	popularity <dbl>
1	Inception	8.364	34495	2010	825532764	FALSE	160000000	83.952
2	Interstellar	8.417	32571	2014	701729206	FALSE	165000000	140.241
3	The Dark Knight	8.512	30619	2008	1004558444	FALSE	185000000	130.643
4	Avatar	7.573	29815	2009	2923706026	FALSE	237000000	79.932
5	The Avengers	7.710	29166	2012	1518815515	FALSE	220000000	98.082
6	Deadpool	7.606	28894	2016	783100000	FALSE	58000000	72.735

6 rows | 1-9 of 12 columns

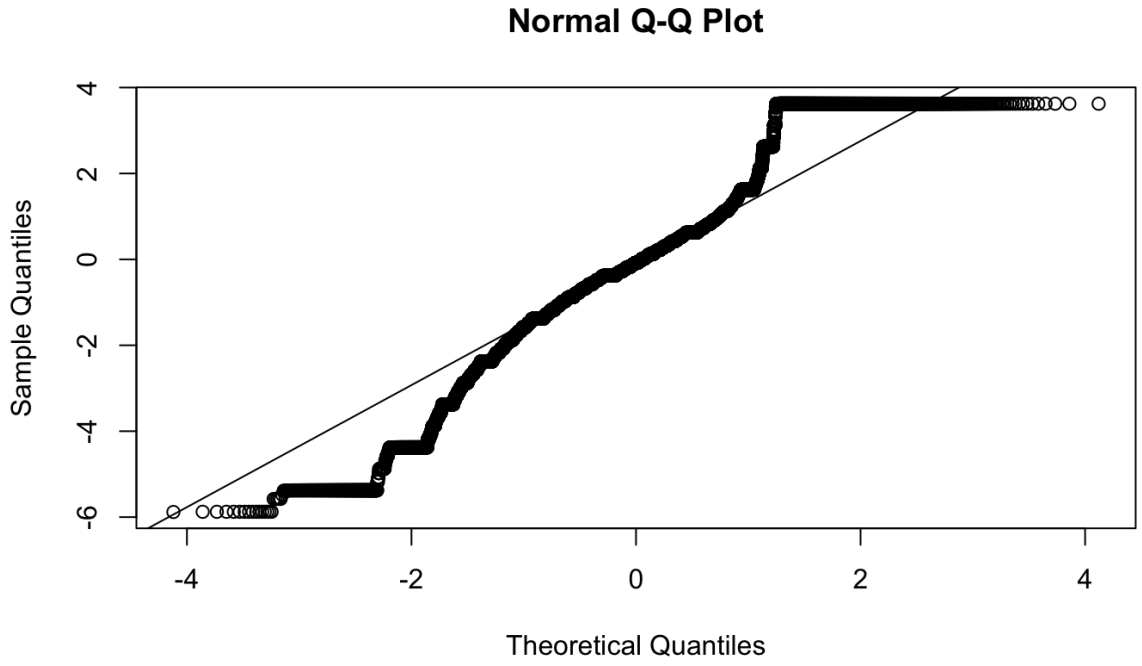
95% Confidence Intervals:
vote_average: (5.993463 6.006672)
vote_count: (63.45697 67.51040)
revenue: (41652458 45523986)
budget: (10002848 10637007)
popularity: (2.636532 2.726043)

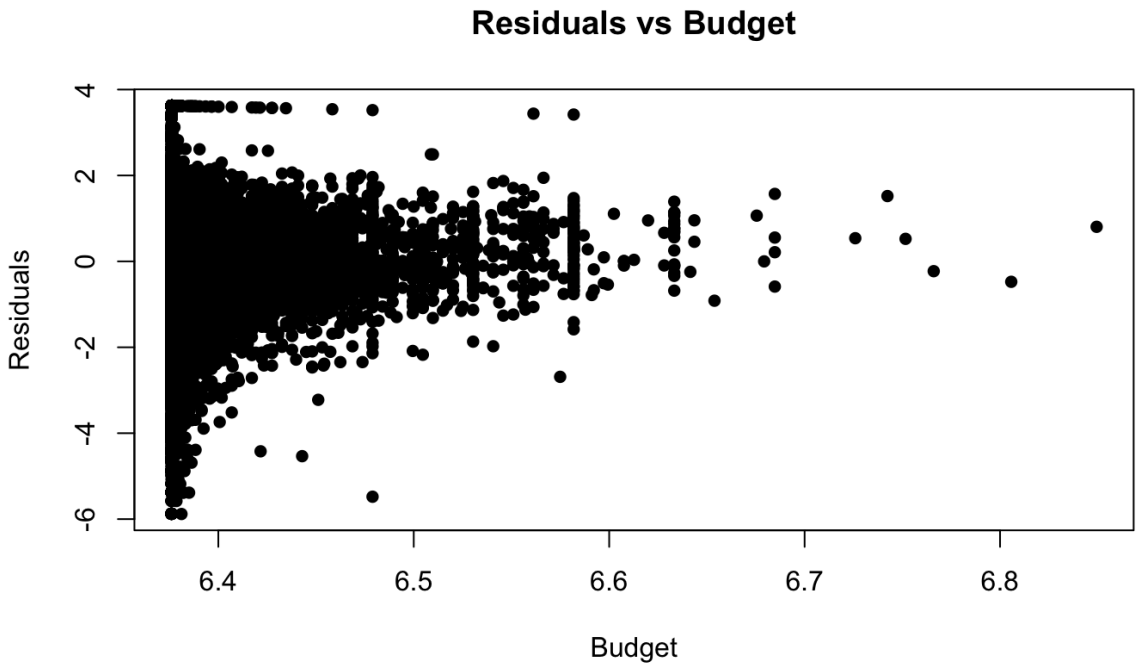
2021 average movie revenue in the U.S is $10,165,556 (https://www.boxofficemojo.com/year/)

Graphs:

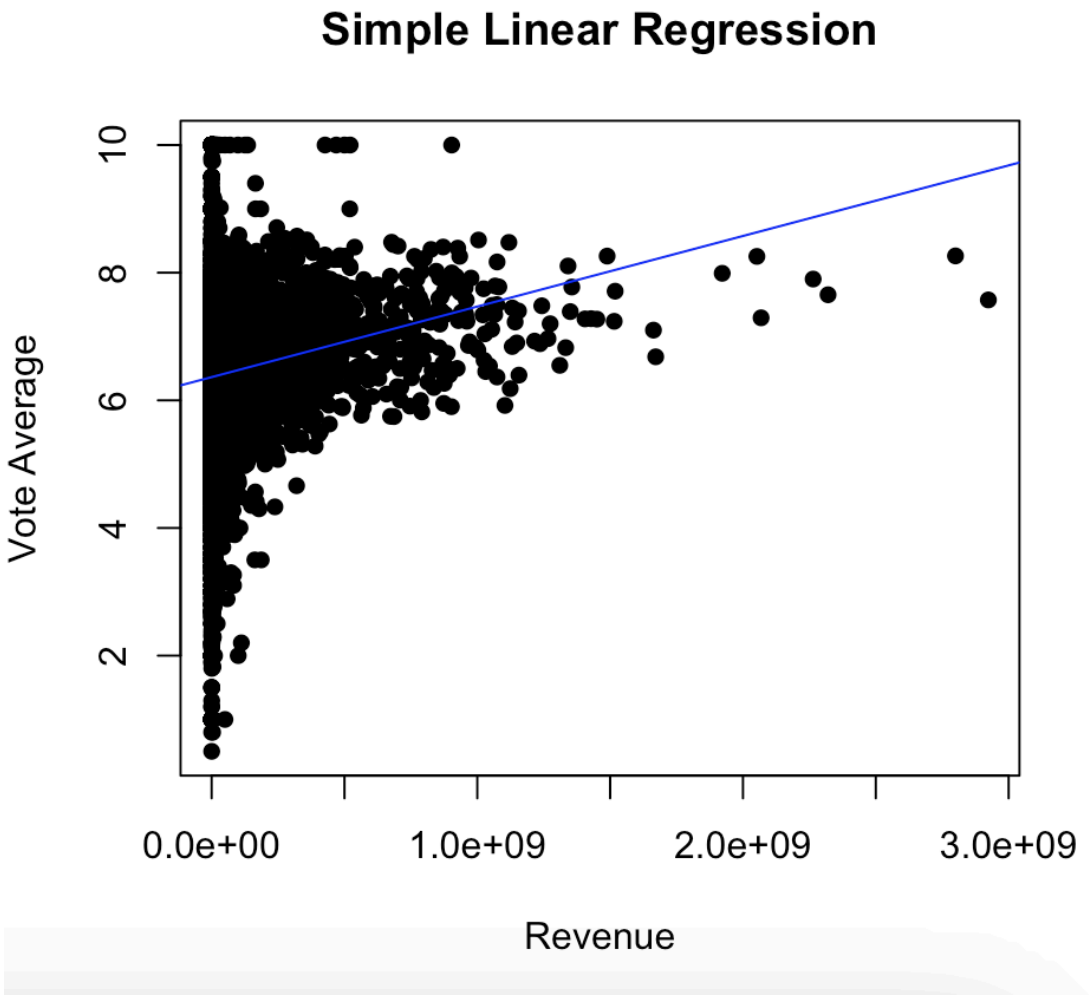


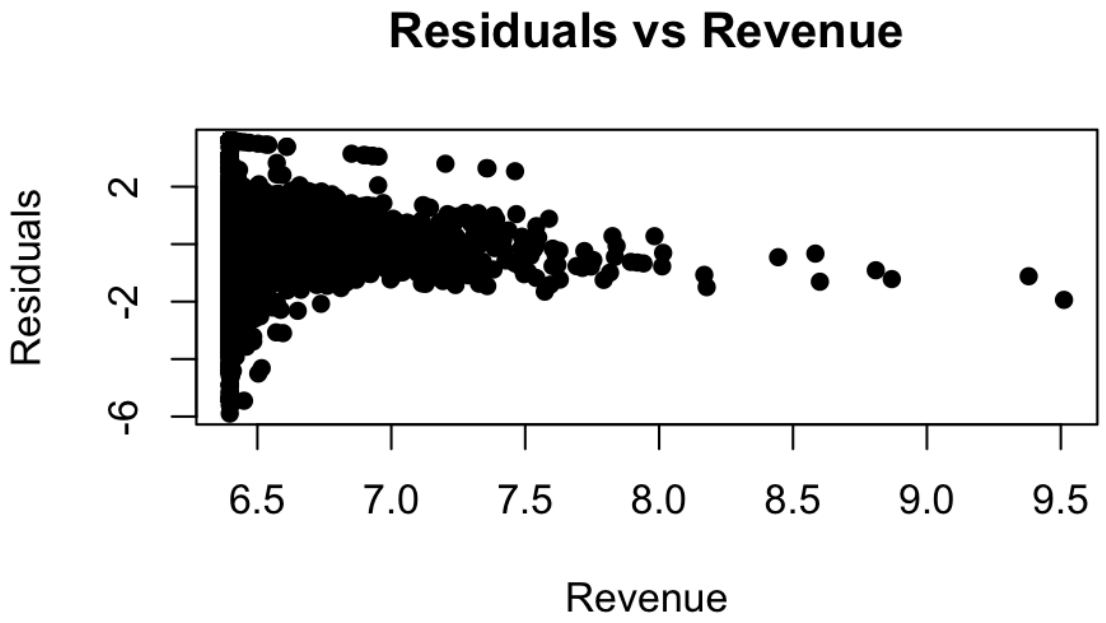
Interpretation: the regression model seems to have a very weak positive correlation. The relationship seems to be insignificant.

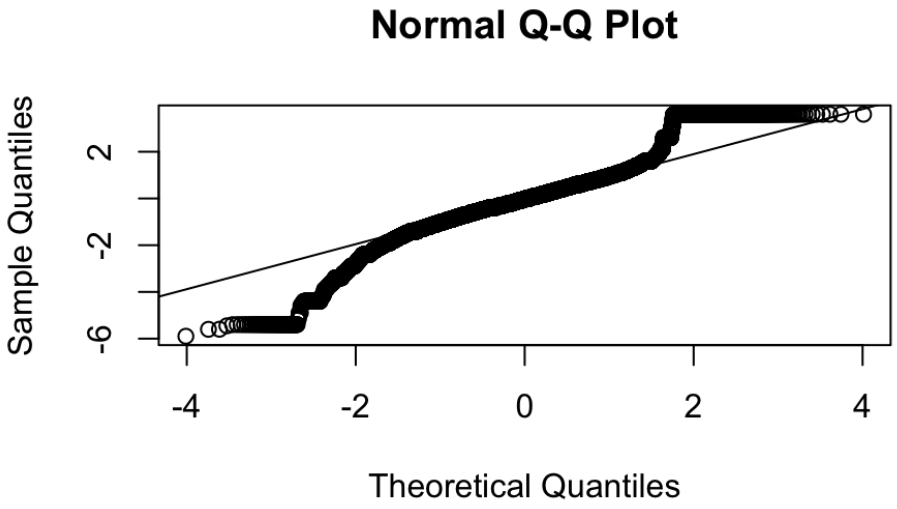




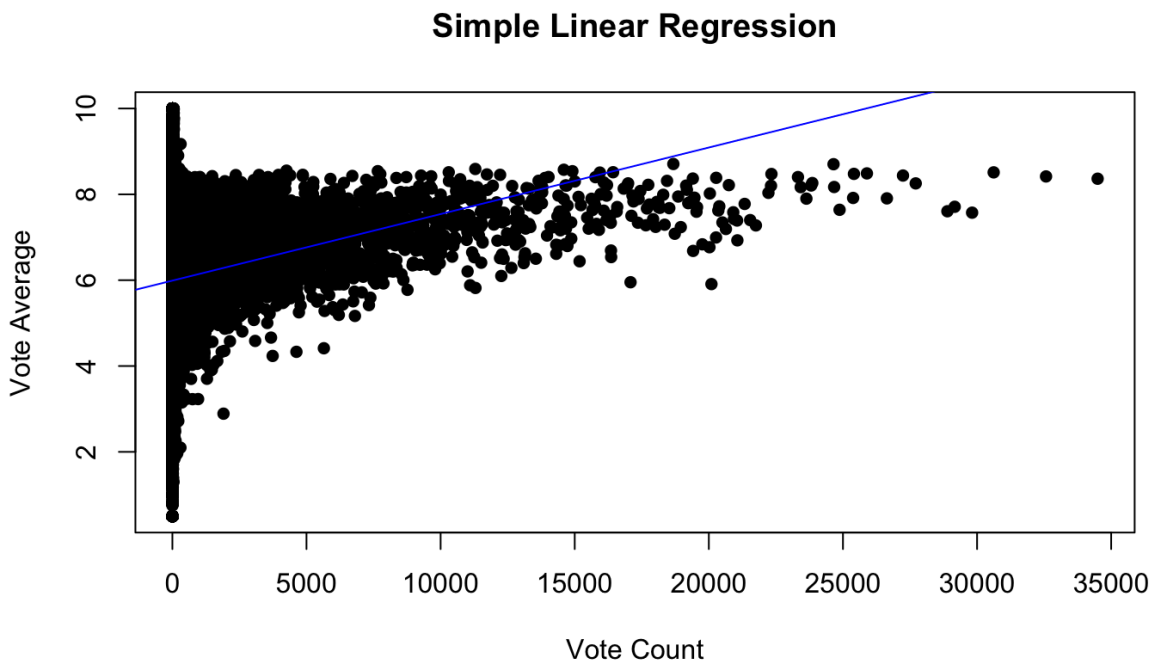
Interpretation: the residual plot doesn't seem to be randomly distributed around 0, and there is a cluster of values when budget is at one end of the range. So budget would not be a good predictor of the vote_average and we need to transform it using square root.



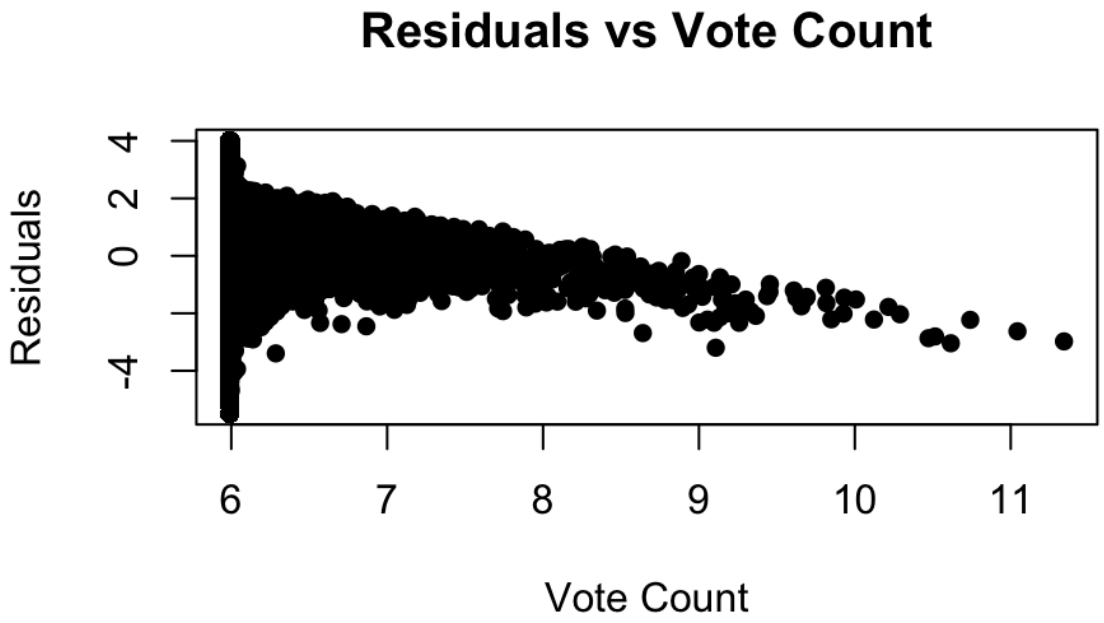




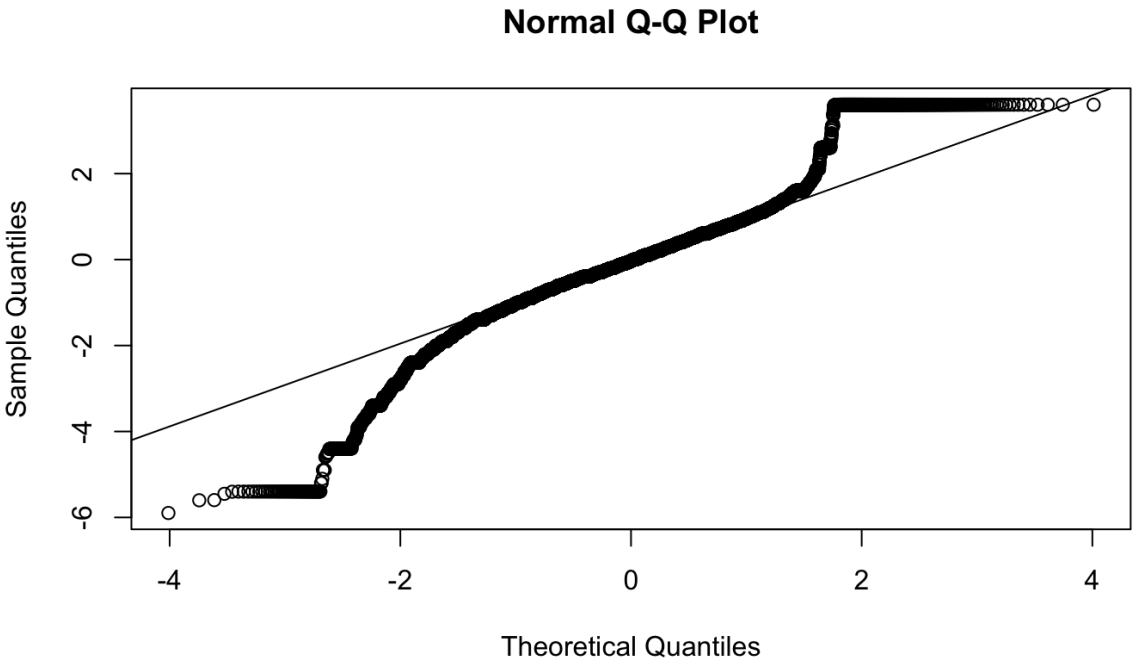
Interpretation: the residual plot and the qq-plot both violate the normality assumption. The residual plot shows more data points are predicted at the lower range, which doesn't seem to be normal. As discussed previously, we may need to take the square root term on the explanatory variable to try to fix the problem around 6.5.



Interpretation: first the linear regression model violates the normality assumption for each value of vote count. Second, there seems to be no relationship between the two variables when vote count gets large.

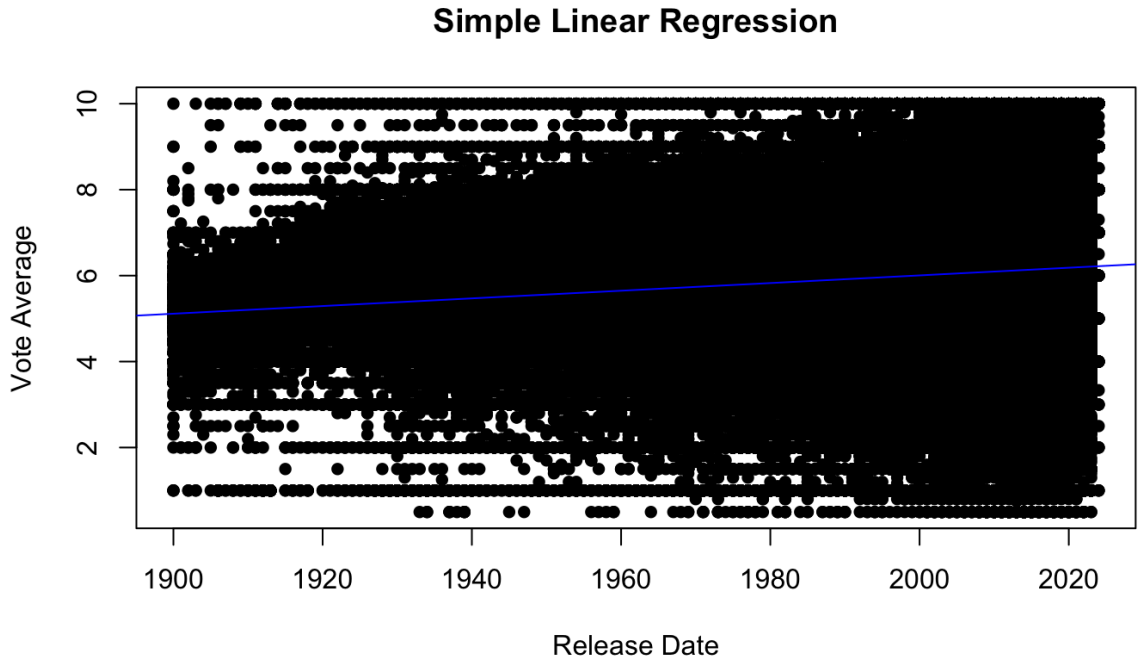


Assumption: the residual plot doesn't look random.



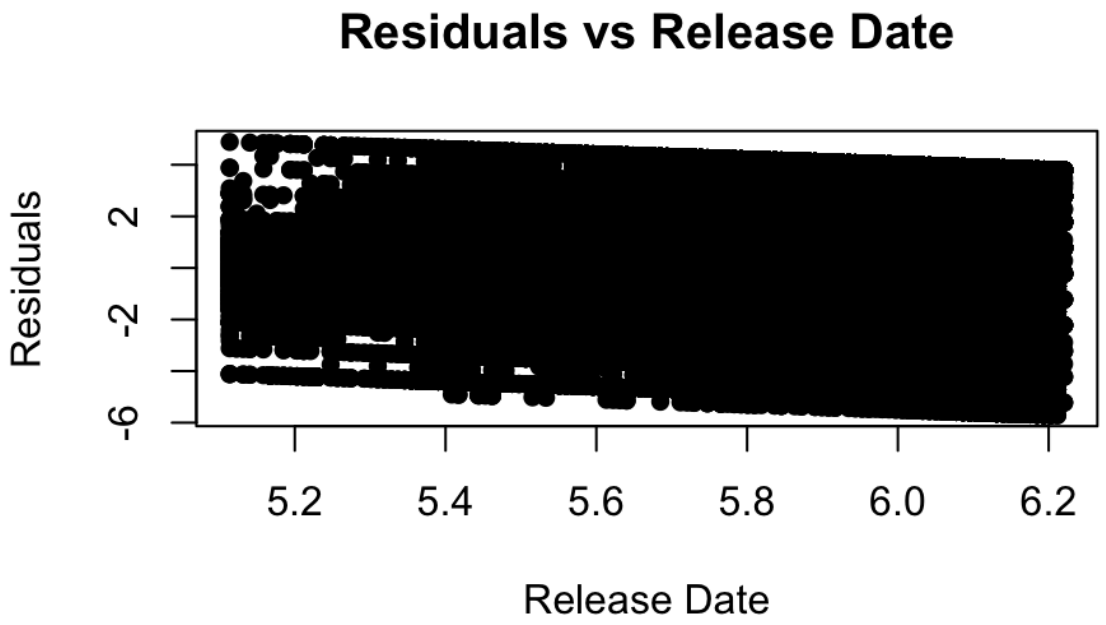
Interpretation: the QQ-plot does not sit on the line so the residuals are not normally distributed. In addition, there seems to be a deviation from the line for the two end, so the residual may be

left-skewed or right-skewed, so we will try to take Log on both variables and try with this variable.

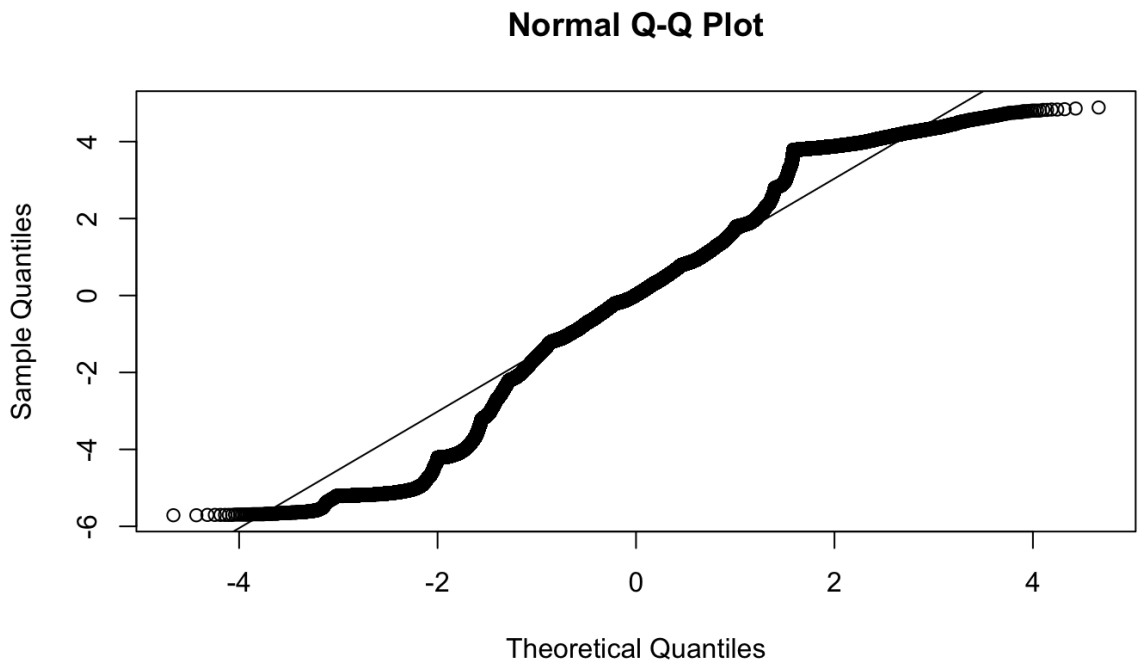


Interpretation: While there is a very slight upward trend in the Vote Average over time, this trend is minimal, and the Release Date does not appear to be a strong predictor of the Vote

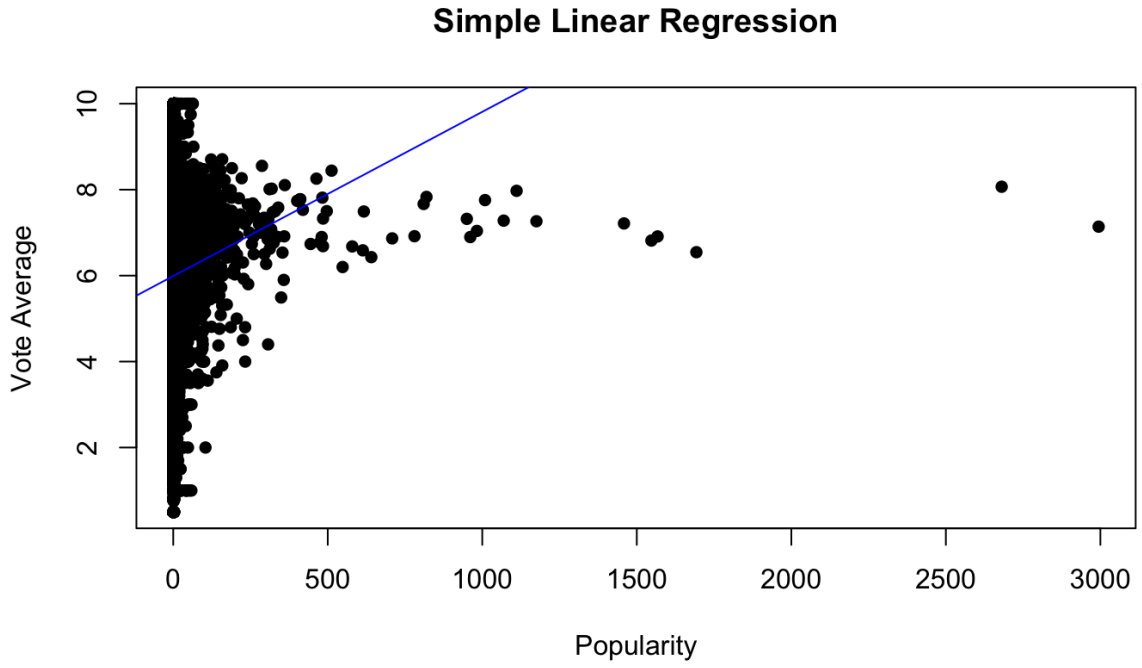
Average.



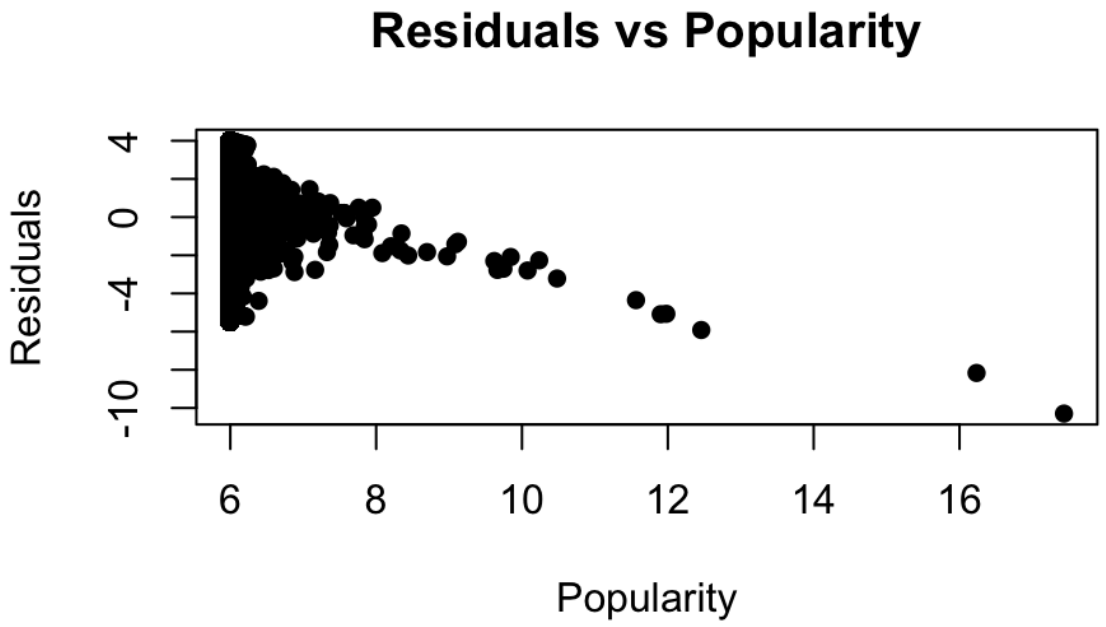
Interpretation: The model's residuals for different predicted "Release Dates" are relatively randomly distributed, with no obvious patterns or systematic bias associated with release dates.



Interpretation: The residuals do not strictly follow a normal distribution due to noticeable deviations in both tails.



Interpretation: while there is a positive trend between popularity and the vote average, the data's spread and concentration patterns suggest potential limitations in a linear model for accurately predicting scores across the full popularity range.



Interpretation: The residuals are clustered near zero for lower values of predicted Popularity but tend to show a wider spread as predicted Popularity increases. The spread of residuals increases with Popularity, indicating a potential issue with heteroscedasticity. Some extreme residual values exist for high Popularity data points (above 2000).

Check assumptions: Based on the qq-plot and histogram of residuals, the normality assumption seems to be violated as shape of the histogram is not approximately asymmetrical and the frequency of residuals at one end is large. Plus, the residuals don't seem to be fitted on the 45 degree line in qq-plot. This suggests the model is not adequate, and our predictor (budget) or explanatory var vote_Average or both is/are not normally distributed, and hence we may need to adjust it by taking Log() on one or both of them. Overall, since the residual data points on the qq-plot and histogram do seem to exhibit extreme left or right skewness, we will try to transform it to be a more normal distribution, and then decide whether to include or not.

Problem: Many questions arose throughout the research process thus far, most of which deal with the complex/abnormal nature of our data. We found many questions that may be answered by transforming the data, like what to do with outliers and deciding if some of our entries are incorrect/misleading and need to be thrown out. All in all, we are trying to account for the intricacies of our dataset while still providing an accurate analysis, which can be confusing and potentially not linear.

Transformations: We plan to transform the data to the format y - a = (x -b)^½ because our data seems to show a sort of square root distribution, and we will experiment with this in the future and reanalyze our model and data, including the outliers. We may see the outliers now no longer being outliers within our transformed model/data so we will analyze and decide to drop or keep them as necessary.

Milestone 4

Introduction

Understanding the factors that influence movie success has become increasingly important in today's era when the movie industry has become a major cultural and economic force. As the number of movies released each year continues to grow, filmmakers and companies are increasingly interested in predicting which movies will resonate with audiences. The purpose of this project is to explore the relationship between a variety of factors (e.g., genre, budget, and director) and a movie's average rating. We aim to analyze qualitative factors including genre and production company, as well as quantitative variables such as budget and revenue, in order to reveal key predictors of a film's success. Despite the fact that the dataset used is less credible in terms of source and collection methodology, we will utilize a comprehensive dataset from Kaggle that seeks to identify significant predictive variables that influence movie success.

Our analysis will focus on both qualitative and quantitative factors. We hypothesize that there may be a positive correlation between higher budgets and revenues and higher average ratings, as these factors typically imply higher production quality and stronger marketing. In addition, we will examine whether specific genres of movies, particularly action movies, receive higher ratings because they cater to audiences' need for thrill-seeking experiences. The role of directors will also be included in the study, as their unique narrative techniques may significantly affect audience engagement and emotional response. Also, we will explore how language and release date affect ratings, taking into account the significant impact of cultural identity and timing choices on audience ratings.

By articulating these hypotheses, we hope to contribute to an understanding of what factors make a movie successful in the eyes of the audience. Through this study, we expect to reveal the intricate relationships between these variables and provide insights that will help filmmakers make more informed decisions that will enhance the potential for movie success.

In our analysis, we will explore both qualitative and quantitative factors to uncover key predictors of a movie's success, as measured by average rating scores. The dependent variable in our study is the average rating of the movie, which ranges from 0 to 10, serving as a response variable that can be influenced by various independent (predictor) variables. Among these predictor variables, genre is a qualitative factor that reflects the diverse audience preferences and expectations; for example, action movies often receive higher ratings from fans who enjoy thrilling and fast-paced narratives. Popularity, measured quantitatively, is another important factor, as popular movies typically generate more buzz, leading to increased viewership and reviews that can positively impact ratings.

Budget, also a quantitative variable, is assumed to correlate with higher average ratings due to the potential for better production values and the inclusion of well-known actors. The production company is a qualitative variable, with certain companies recognized for their successful films at festivals; we hypothesize that movies produced by more reputable companies will achieve higher ratings. Additionally, the director's influence is crucial, as their unique visual style and storytelling approach can significantly affect the audience's emotional response, thus impacting ratings. The spoken language of a film plays a qualitative role as well; movies in familiar languages may resonate better with audiences, enhancing their viewing experience and ratings. The release date, another qualitative factor, can influence ratings, as films released during peak seasons, like holidays or summer, often attract larger audiences. Furthermore, revenue serves as a quantitative indicator of a movie's appeal; high revenue typically reflects strong audience interest, which may translate into more positive reviews. Lastly, the production countries—indicating where the film was originally produced—may also affect audience preferences, potentially leading to higher ratings for films from countries that viewers favor, such as American movies.

In our analysis, we aim to explore various relationships between the predictor variables and movie ratings within our dataset. We propose several preliminary hypotheses that guide our investigation. First, we hypothesize that higher revenue is associated with higher movie ratings, as substantial financial returns often indicate strong audience interest and positive reception. Second, we believe that a higher budget will similarly predict elevated movie ratings, given that increased funding can enhance production quality and attract prominent actors. Additionally, we hypothesize that specific genres, particularly action films, will generally receive higher average ratings due to their appeal to audiences seeking thrilling experiences. Finally, we anticipate that greater popularity, measured by factors such as audience buzz and viewership, will also correlate with higher movie ratings. By examining these hypotheses, we hope to gain insights into the key factors that contribute to a film's success.

We cleaned the data by removing observations with a value of 0 and retaining only the independent variables and dependent variables we wish to explore. We used R to perform graphical analysis, including creating Q-Q plots and residuals versus variable plots, to visually assess the normality of the data and the linear relationships. We interpreted these plots for each variable, determined if any transformations are necessary, and evaluated whether any variables should be dropped. Then, we identified outliers in the data and decided on appropriate actions to address them. We found that model assumptions are all violated, which leads us to try more interactions and transformation on variables in milestone 4. Last, we performed multicollinearity check, model fit, and variable selections to conclude our final model.

Full model

Initially, we incorporate five basic quantitative data in the model, which are budget, vote_count, popularity, release_date, and revenue.

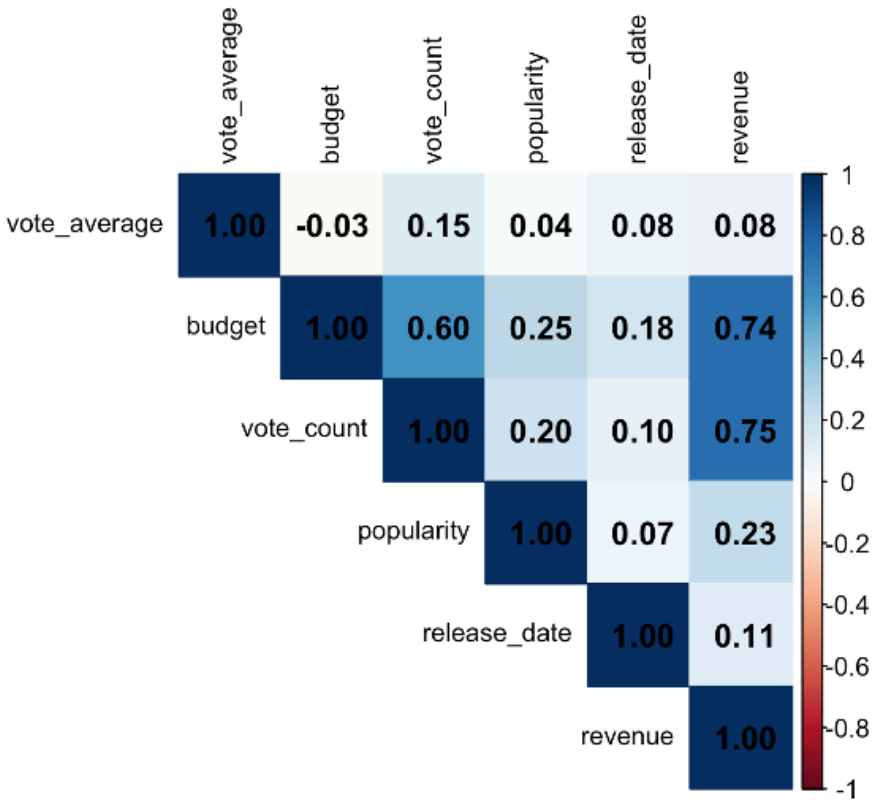
```
Call:
lm(formula = vote_average ~ budget + vote_count + popularity +
    release_date + revenue, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7717 -0.6355 -0.0431  0.5436  4.6600

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.097e+00  1.307e+00  -3.900 9.66e-05 ***
budget      -7.534e-09  4.884e-10 -15.424  < 2e-16 ***
vote_count   1.084e-04  6.244e-06  17.360  < 2e-16 ***
popularity   7.430e-04  2.064e-04   3.599 0.000321 ***
release_date 5.805e-03  6.534e-04   8.883  < 2e-16 ***
revenue      4.374e-10  1.432e-10   3.055 0.002255 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.231 on 10538 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.05848,   Adjusted R-squared:  0.05804
F-statistic: 130.9 on 5 and 10538 DF,  p-value: < 2.2e-16
```

From the summary of the full model, we can see that all the variables are statistically significant. However, the model has a very low adjusted R2 square, indicating that only 5% of the variation in average vote scores is explained by the model.

Multicollinearity

In order to avoid redundancies in the model, we want to check for the multicollinearity by generating quantitative covariates to see if there are correlations between predictors.

After using pairwise comparison, we can see that budge, revenue, and vote_count all have two highly correlated covariates.



To determine which variables should be removed, we removed each variable one at a time and checked the adjusted R square after each change.

Removing vote_count resulted in a decrease in the adjusted R square:

```
Residual standard error: 1.248 on 10539 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.03156,   Adjusted R-squared:  0.03119
F-statistic: 85.86 on 4 and 10539 DF,  p-value: < 2.2e-16
```

.

Removing budget resulted in a decrease in the adjusted R square:

```
Residual standard error: 1.245 on 10539 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.03723,   Adjusted R-squared:  0.03686
F-statistic: 101.9 on 4 and 10539 DF,  p-value: < 2.2e-16
```

Removing revenue resulted in little to no change in the adjusted R square:

```
Residual standard error: 1.232 on 10539 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.05765,   Adjusted R-squared:  0.05729
F-statistic: 161.2 on 4 and 10539 DF,  p-value: < 2.2e-16
```

Therefore, we decided to include vote_count, popularity, release_date, and budget as covariates in the model for this step.

Model fit

```
Call:
lm(formula = vote_average ~ vote_count + popularity + release_date +
    budget, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7644 -0.6359 -0.0417  0.5430  4.7102

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.980e+00  1.307e+00  -3.811 0.000139 ***
vote_count    1.193e-04  5.133e-06  23.236  < 2e-16 ***
popularity    7.750e-04  2.063e-04   3.758 0.000173 ***
release_date  5.743e-03  6.534e-04   8.789  < 2e-16 ***
budget       -6.718e-09  4.091e-10 -16.421  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.232 on 10539 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.05765,   Adjusted R-squared:  0.05729
F-statistic: 161.2 on 4 and 10539 DF,  p-value: < 2.2e-16
```

The full model includes the covariates: vote_count, popularity, release_date, and budget.

- **F-test:**
  - F-statistic: 161.2 and p-value: $<2.2e{-}16$ which is really small.
  - Conclusion: Since the p-value is much smaller than 0.05, the model as a whole is statistically significant, meaning at least one predictor is associated with the outcome variable, vote_average.
- **Individual Predictors:**
  - All predictors (vote_count, popularity, release_date, budget) are statistically significant ($p<0.001$), indicating that they contribute to the model.

- **Multiple R-square:** 0.05765 & **Adjusted R-square:** 0.05729

These values are very low, meaning that the model explains only about 5.7% of the variance in the response variable (vote_average). While the model is statistically significant, its practical explanatory power is weak.

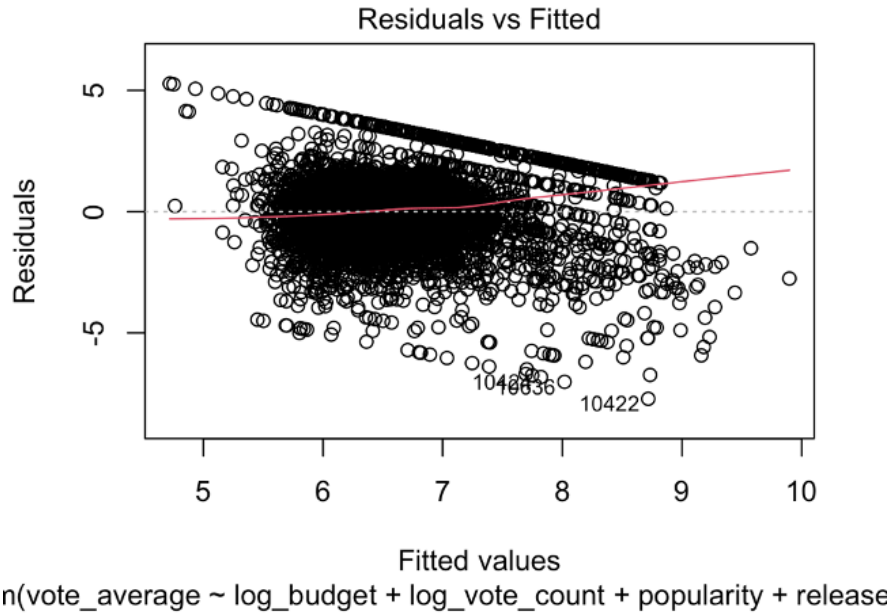
The low R square suggests that there may be a need for non-linear transformations to better capture the relationship between the predictors and the response variable. Therefore, we try some transformation below:

- **Log-transforming skewed predictors budget and vote_count.**

```
Call:
lm(formula = vote_average ~ log_budget + log_vote_count + popularity +
    release_date, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7179 -0.5476  0.0513  0.6166  5.2799

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.9720054  1.2044792   0.807     0.42
log_budget     -0.2164385  0.0045726 -47.334  < 2e-16 ***
log_vote_count  0.2014658  0.0067736  29.743  < 2e-16 ***
popularity      0.0012614  0.0001903   6.627 3.58e-11 ***
release_date    0.0038451  0.0006001   6.407 1.54e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.145 on 10539 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.1849,    Adjusted R-squared:  0.1845
F-statistic: 597.5 on 4 and 10539 DF,  p-value: < 2.2e-16
```

Adjusted R squared is 0.1845, which indicates that the model explains approximately 18.45% of the variance in the response variable (vote_average). This is a significant improvement over the previous model without transformations (R squared ≈5.7%).

F-statistic is 597.5 and $p<2.2\times10-16$, suggesting that the model as a whole is statistically significant.



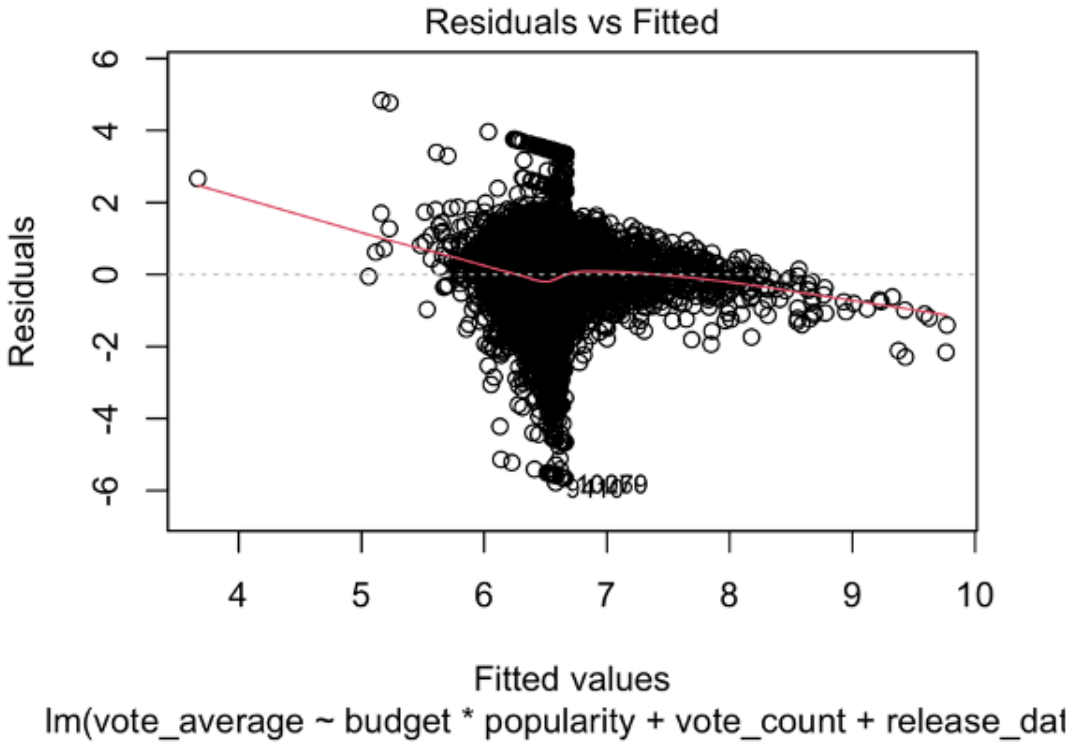
n(vote_average ~ log_budget + log_vote_count + popularity + release

The residuals are centered around zero, suggesting no severe bias in predictions. There is some curvature in the residuals, indicating possible non-linearity not fully addressed by the transformations. A slight funnel shape suggests heteroscedasticity, where residual variance increases for higher fitted values.

- **Including interaction terms: budget * popularity**

```
Call:
lm(formula = vote_average ~ budget * popularity + vote_count +
    release_date, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7827 -0.6366 -0.0353  0.5448  4.8372

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.036e+00  1.305e+00  -3.858 0.000115 ***
budget            -7.468e-09  4.359e-10 -17.132  < 2e-16 ***
popularity        -3.890e-04  3.129e-04  -1.243 0.213825
vote_count         1.229e-04  5.179e-06  23.727  < 2e-16 ***
release_date       5.781e-03  6.527e-04   8.857  < 2e-16 ***
budget:popularity  1.310e-11  2.651e-12   4.944 7.78e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.23 on 10538 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.05983,   Adjusted R-squared:  0.05939
F-statistic: 134.1 on 5 and 10538 DF,  p-value: < 2.2e-16
```

The adjusted R squared indicates that the model explains approximately 5.94% of the variance in vote_average. While this is slightly higher than the baseline model, the explanatory power remains weak. The F-statistic is 134.1 and $p < 2.2 \times 10^{-16}$, which indicates that the model as a whole is statistically significant.



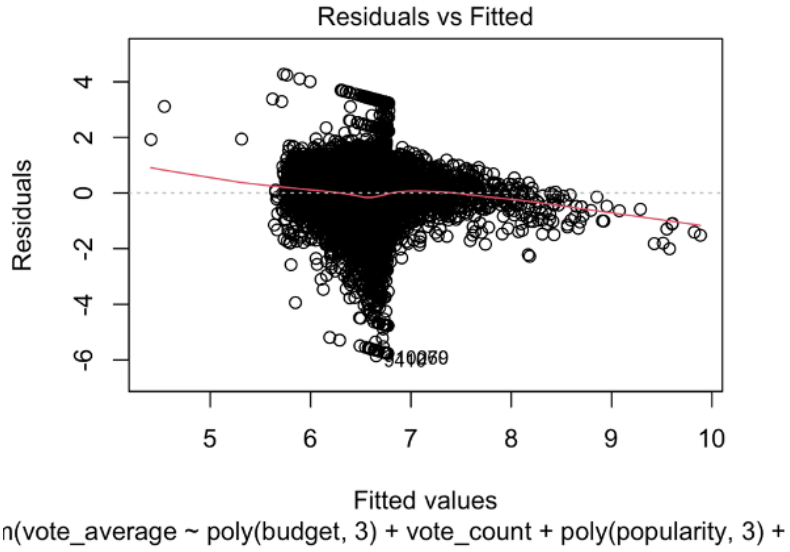
There is some curvature in the residuals, suggesting that the model may not fully capture non-linear relationships. There is a slight funnel shape, indicating heteroscedasticity (non-constant variance of residuals).

While the inclusion of the interaction term improves the model's fit slightly (adjusted R squared increased to 5.94%), the improvement is marginal. The interaction term provides additional insight into how budget and popularity jointly influence vote_average, but the overall explanatory power remains low.

- **Checking for Cubic Polynomial Transformation**

```
Call:
lm(formula = vote_average ~ poly(budget, 3) + vote_count + poly(popularity,
    3) + release_date, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.8536 -0.6223 -0.0105  0.5534  4.2688

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)         -6.393e+00  1.299e+00  -4.920 8.77e-07 ***
poly(budget, 3)1    -2.759e+01  1.615e+00 -17.084  < 2e-16 ***
poly(budget, 3)2     1.414e+01  1.233e+00  11.467  < 2e-16 ***
poly(budget, 3)3    -1.052e+01  1.235e+00  -8.520  < 2e-16 ***
vote_count           1.160e-04  5.475e-06  21.193  < 2e-16 ***
poly(popularity, 3)1  5.503e+00  1.284e+00   4.287 1.83e-05 ***
poly(popularity, 3)2 -3.138e+00  1.358e+00  -2.311   0.0209 *
poly(popularity, 3)3  2.154e+00  1.379e+00   1.562   0.1183
release_date         6.385e-03  6.490e-04   9.838  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.22 on 10535 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.07569,   Adjusted R-squared:  0.07498
F-statistic: 107.8 on 8 and 10535 DF,  p-value: < 2.2e-16
```

The model explains approximately 7.5% of the variance in vote_average, which is slightly higher than the earlier interaction model (5.9%). The f-statistic is 107.8 and p <2 .2×10−26, which means the overall model is statistically significant**.**



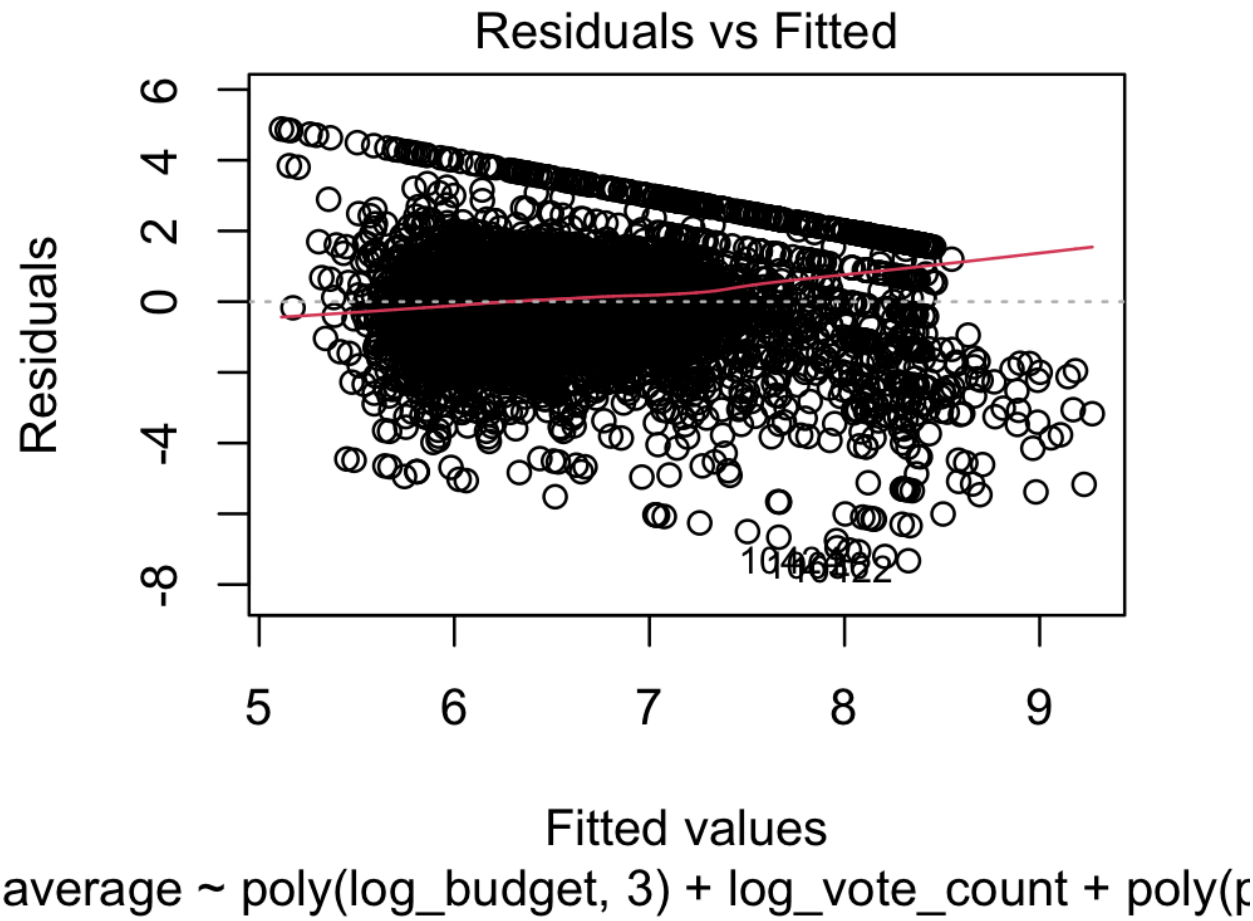
The residuals are centered around zero, and the red line is relatively flat, suggesting that the model has captured most of the non-linearity. There is still some funnel shape, indicating potential heteroscedasticity (residuals spread increases for higher fitted values).

The cubic model (7.5%) explains more variance than both the interaction model (5.9%) and the original model (5.7%). The residual plot shows improvement in capturing non-linear relationships, particularly for budget and popularity.

- **Checking for Log and Cubic Polynomial Transformation**

```
> summary(model_polylog)

Call:
lm(formula = vote_average ~ poly(log_budget, 3) + log_vote_count + 
    poly(popularity, 3) + release_date, data = cleaned_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.3272 -0.5342  0.0570  0.6295  4.8857 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           1.428e+00  1.263e+00   1.131 0.258187    
poly(log_budget, 3)1 -8.274e+01  1.839e+00 -44.999  < 2e-16 ***
poly(log_budget, 3)2  3.500e+00  1.359e+00   2.575 0.010035 *  
poly(log_budget, 3)3  1.149e+01  1.232e+00   9.330  < 2e-16 ***
log_vote_count        1.661e-01  7.841e-03  21.188  < 2e-16 ***
poly(popularity, 3)1  8.102e+00  1.191e+00   6.804 1.07e-11 ***
poly(popularity, 3)2 -8.528e+00  1.258e+00  -6.781 1.26e-11 ***
poly(popularity, 3)3  7.922e+00  1.286e+00   6.159 7.59e-10 ***
release_date          2.127e-03  6.283e-04   3.386 0.000713 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.135 on 10535 degrees of freedom
  (189 observations deleted due to missingness)
Multiple R-squared:  0.2004,    Adjusted R-squared:  0.1998 
F-statistic: 330.1 on 8 and 10535 DF,  p-value: < 2.2e-16
```

The adjusted R^2 has increased to 0.1998, explaining approximately 19.98% of the variance in vote_average. This is an improvement over previous models, suggesting the inclusion of polynomial terms better captures the non-linear relationships in the data.



The residuals are centered around zero, with no severe patterns. The residual plot shows some curvature and a slight funnel shape, indicating potential non-linear relationships and heteroscedasticity that are not fully addressed by the polynomial terms.

Since the transformations did not lead to a significant improvement in the model's performance, we have decided to retain the original full model for further analysis.

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

Residuals:
    Min      1Q  Median      3Q     Max
-48.266  -4.217   0.992   2.156   6.912

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.207e+00  6.345e-03 663.000   <2e-16 ***
vote_count   4.940e-04  1.400e-05  35.295   <2e-16 ***
popularity   1.680e-02  4.177e-04  40.205   <2e-16 ***
revenue     -5.190e-09  3.122e-10 -16.620   <2e-16 ***
budget       8.734e-09  9.511e-10   9.183   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.924 on 224155 degrees of freedom
Multiple R-squared:  0.02093,   Adjusted R-squared:  0.02091
F-statistic:  1198 on 4 and 224155 DF,  p-value: < 2.2e-16
```

The linear regression model indicates that vote_count, popularity, and budget have positive coefficients, suggesting a positive correlation with vote_average (movie ratings), while revenue has a negative coefficient, indicating a potential negative correlation. All predictors are statistically significant, with p-values less than 0.001. However, the model's adjusted R^2 is only 0.02091, meaning the model explains just 2.091% of the variance in movie ratings, demonstrating weak explanatory power. The residual standard error is 2.924, indicating an average prediction error of approximately 2.924 rating points. Although the model is statistically significant (p-value of the F-test < 2.2e-16), its practical ability to predict ratings is limited.

Variable selection

Because of those small R^2 values, I tried some other methods of scaling and regression called lasso regression and using the scaled method. With the lasso regression, which ____, on the unscaled data (cleaned_data), I found these coefficients which were quite small and imply almost no influence on vote average which means our model is very poor and our variables have almost no effect on our response variable. Then I tried it on the scaled data and found better coefficients implying more of an influence, but, while these values appear bigger and are easier to interpret, they need to be interpreted in the following context:

During the process of variable selection for our final model, I attempted both the forward and backward selection process and found limited success, meaning our R^2 values for both processes were very, very low (around .057), but both process agreed on the same selection, which is all the variables we included (budget, vote count, popularity, release date). For the forward process, we had the following output:

Variables Selected:
=> vote_count
=> budget
=> release_date
=> popularity

Stepwise Summary

Step	Variable	AIC	SBC	SBIC	R2	Adj. R2
0	Base Model	34802.046	34816.568	4941.782	0.00000	0.00000
1	vote_count	34483.100	34504.884	4622.843	0.03004	0.02995
2	budget	34268.078	34297.123	4407.888	0.04984	0.04966
3	release_date	34196.208	34232.514	4336.058	0.05649	0.05622
4	popularity	34183.497	34227.065	4323.360	0.05781	0.05745

Final Model Output

Model Summary

R	0.240	RMSE	1.227
R-Squared	0.058	MSE	1.506
Adj. R-Squared	0.057	Coef. Var	18.742
Pred R-Squared	0.057	AIC	34183.497
MAE	0.858	SBC	34227.065

RMSE: Root Mean Square Error
MSE: Mean Square Error
MAE: Mean Absolute Error
AIC: Akaike Information Criteria
SBC: Schwarz Bayesian Criteria

ANOVA

	Sum of Squares	DF	Mean Square	F	Sig.
Regression	972.523	4	243.131	161.318	0.0000
Residual	15850.722	10517	1.507		
Total	16823.245	10521			

Parameter Estimates

model	Beta	Std. Error	Std. Beta	t	Sig	lower	upper
(Intercept)	-4.594	1.304		-3.523	0.000	-7.150	-2.038
vote_count	0.000	0.000	0.276	23.334	0.000	0.000	0.000
budget	0.000	0.000	-0.198	-16.374	0.000	0.000	0.000
release_date	0.006	0.001	0.082	8.510	0.000	0.004	0.007
popularity	0.001	0.000	0.038	3.836	0.000	0.000	0.001

And for the backwards selection process we got the following output:

Step => 0
Model => vote_average ~ vote_count + popularity + release_date + budget
R2 => 0.058

No more variables to be removed.

Because of our low results, I researched some other methods of variable selection and scaled regression and concluded that using Lasso regression was a great option because it scales the coefficients of the model towards zero and eliminates informative variables in that process. Unfortunately, as our variables are all very uninformative, the lasso model also returned the same coefficients, meaning there wasn't much to any multicollinearity present within our variables and scaling wasn't really the main issue. But, to further test the effect of scaling, I used the scale() method on all of our independent variables and ran the lasso once again and I did find different

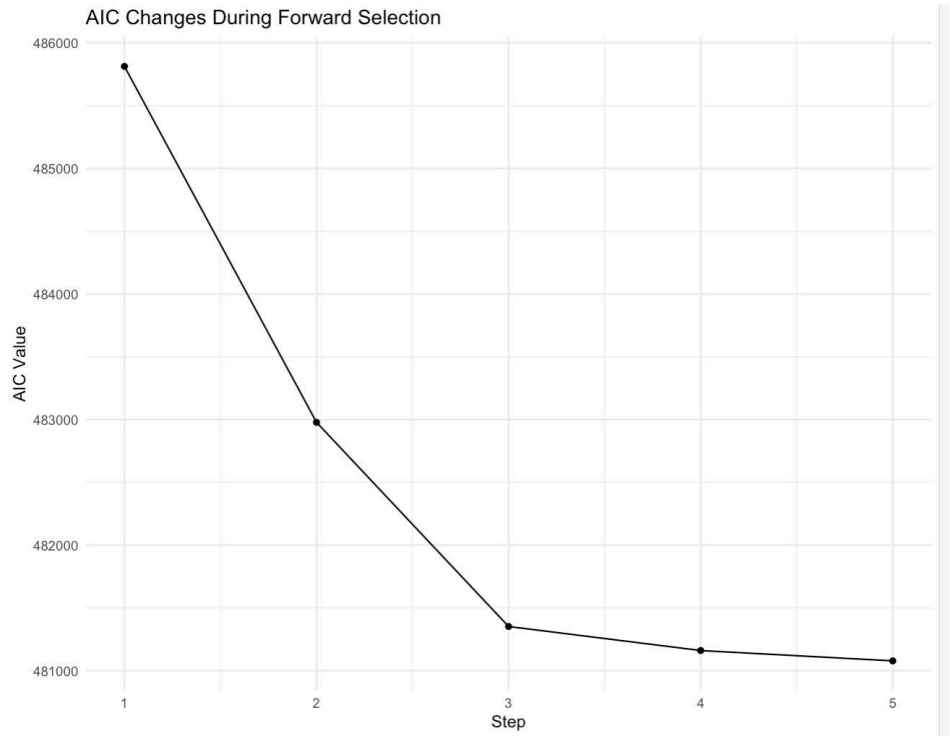
coefficients for the model. The unscaled and scaled coefficients are shown below. While they look more informative, in reality they are scaled differently, so they are just as uninformative and their effects are still negligible. The scaled() function scales the data to be interpreted instead of with one unit increase in each variable we expect an increase of the coefficient for the response, the scaled assumes a one standard deviation increase in each variable expects an increase of the coefficient for the response variable.

Unscaled:
6 x 1 sparse Matrix of class "dgCMatrix"
s0
(Intercept) -4.625731e+00
vote_count 1.085149e-04
revenue 4.187166e-10
budget -7.412917e-09
popularity 7.461933e-04
release_date 5.567536e-03

Scaled:
6 x 1 sparse Matrix of class "dgCMatrix"
s0
(Intercept) -4.602609161
vote_count 0.317210516
revenue 0.064324203
budget -0.278063600
popularity 0.044986788
release_date 0.005571258

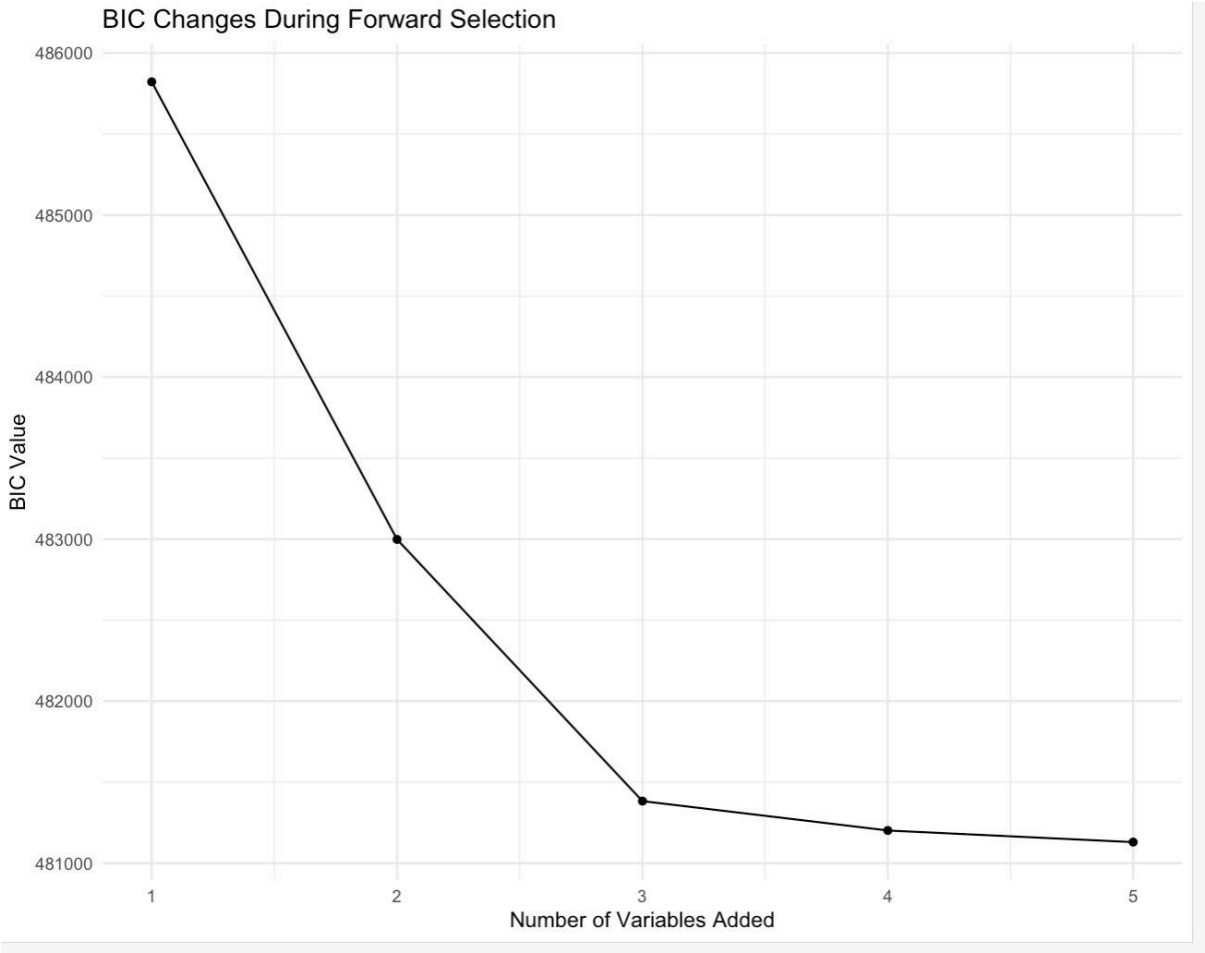
In addition to using the p-value criteria for forward and backward, we tried AIC & BIC based criteria so that capturing and checking variables that should be included in the model.

Forward selection AIC based



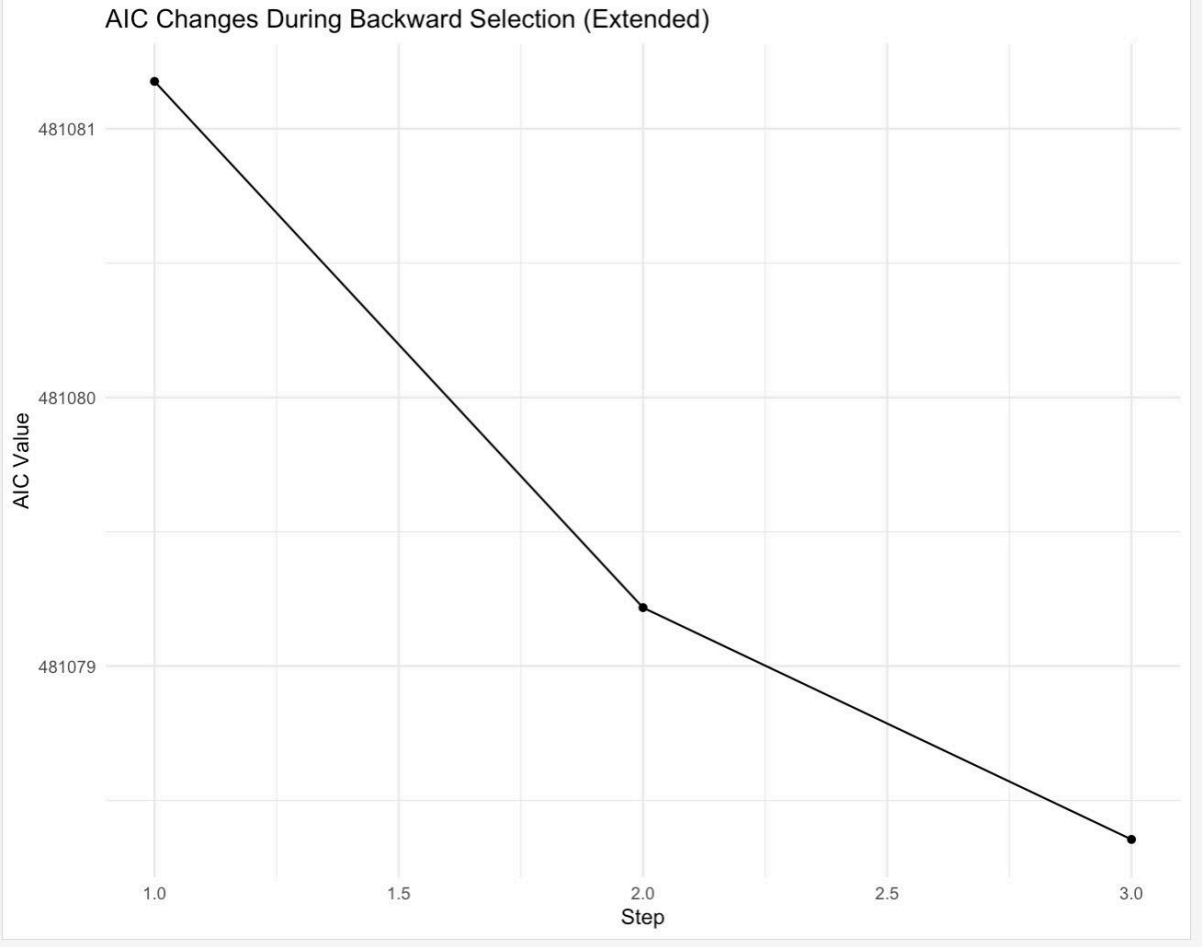
The plot shows the changes in AIC values during the forward selection process, where variables are progressively added to improve the model. The AIC value decreases sharply in the initial steps, particularly from Step 1 to Step 3, indicating that the first few variables significantly enhance the model. After Step 3, the rate of improvement slows, with diminishing returns from adding more variables. The lowest AIC value, around 481,000, is reached at Step 5, suggesting this is the optimal model that balances explanatory power and complexity. This process demonstrates that the most impactful variables are included in the early steps, while later additions have a smaller effect on the model's quality.

Forward selection BIC based



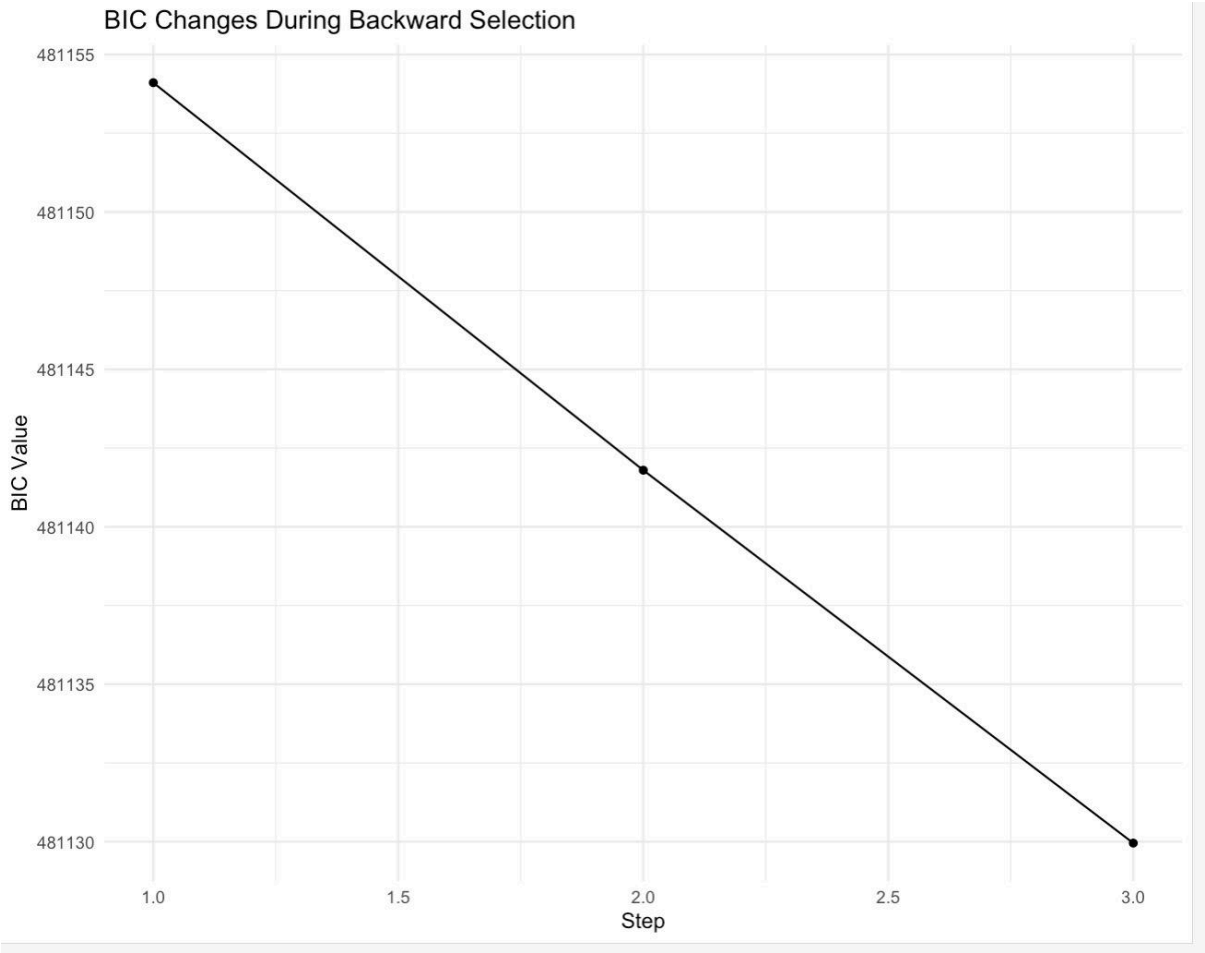
The plot illustrates the changes in BIC during forward selection, showing how the BIC value decreases as more variables are added to the model. Initially, the BIC drops significantly, indicating that the first few variables contribute the most to improving the model's fit. However, from the fourth variable onward, the BIC stabilizes around 481,000, suggesting diminishing returns from adding more predictors. The lowest BIC is achieved at step 5, indicating that the model with five variables strikes the best balance between explanatory power and model complexity. This trend reflects the effectiveness of BIC in penalizing overly complex models while prioritizing simplicity and fit.

Backward selection AIC based



The plot illustrates the changes in AIC values during the backward selection process. Each step represents the removal of one variable from the model, and the AIC value decreases steadily from Step 1 to Step 3, indicating improvements in model quality. The lowest AIC value, approximately 481,079, is achieved at Step 3, suggesting this is the optimal model that balances explanatory power and model complexity. The diminishing decline in AIC values indicates that the initial variable removals had a greater impact on model improvement, while later adjustments contributed less significantly. This process highlights the effectiveness of backward selection in simplifying the model while maintaining its predictive quality.

Backward selection BIC based



The plot shows the changes in BIC (Bayesian Information Criterion) values during the backward selection process. Each step represents the removal of one variable from the model. The BIC value decreases consistently from Step 1 to Step 3, indicating that each variable removed improved the model's balance between fit and complexity. The lowest BIC value, approximately 481,130, is achieved at Step 3, suggesting that the model at this step is the optimal one, as it minimizes BIC. The linear trend in the BIC decrease implies that the variables removed in each step had similar effects on the model's improvement. Overall, the backward selection effectively simplified the model while maintaining or improving its quality.

In summary, all the selection criterias suggests the full model including vote_count, popularity, budget, as well as revenue is more predictive of vote_average, although the estimators are very small, which suggests problems in the initial dataset.

Model interpretation

Our model uses the number of votes, release date, movie revenue, budget, and popularity to predict a movie's performance among the general public, as measured by the average vote score. The coefficients are **-5.097, -0.000000007534, 0.0001084, 0.000743, 0.005805, and 0.0000000004374** for the intercept, budget, vote_count, popularity, release_date, and revenue, respectively.

The intercept doesn't have a natural interpretation, since voting cannot be negative and the predictors cannot be 0 either. The slopes can have natural interpretations if voting is within the range 0-10, and are not highly predictive of variations in voting scores. For example, the release date has a coefficient of 0.005805, indicating that for every one-year increase in the movie's release year, the vote_average score increases by 0.005805, holding other variables constant. This implies a slight improvement in ratings over time. For instance, if a movie from the 1980s has a vote_average of 7, a movie from the 2020s would be expected to have a vote_average of 7.2322, controlling for other factors. However, the difference between scores of 7 and 7.2322 is slight to determine which movie is better-rated, given the range of voting is 10.

To explain what our model is predicting, imagine a random movie that has characteristics as follow:

Budget: $10 million

Vote Count: 5,000

Popularity: 30

Release Date: 1985

Revenue: $50 million

Then, the predicted average voting would be

vote_average =
$-5.097+(-0.000000007534\times10{,}000{,}000)+(0.0001084\times5000)+(0.000743\times30)+(0.005805\times1985)$
$+(0.0000000004374\times50{,}000{,}000)$

which is approximately 6.937.

To evaluate our model's accuracy, we'll apply two recent films *Smile 2* and *Venom: The Last Dance* . For *Smile 2*, with a budget of $28 million, release date in 2024, revenue of $130 million, vote count of 50,759, and popularity of 119, the model predicts a vote_average of 12.089. The 95% confidence interval for the vote_average is (11.991, 12.187), and the 95% prediction

interval for the actual rating is (9.672, 14.506), showing a wider range of intervals. The actual rating is 6.9 according to IMDb, which means that the predicted result is overestimated the actual results.

For *Venom: The Last Dance*, actual rating is 6.2, budget is $12 million, release date is in 2024, revenue is $135,663,186, vote count is 45k, and popularity of 97, according to IMDb. The predicted rating is 11.571 using our model. The 95% Confidence Interval is (11.375, 11.767), and the 95% Prediction Interval is (9.148, 13.994). The actual rating for the movie is 6.2, indicating that our model overestimates the rating for this film.

Overall, our model does not give a good estimation of the movie votings and using recent data in our model would cause overestimation of the result. This might be improved if we incorporate more recent data to predict the model.

Conclusion

In this project, we start by investigating the factors influencing movie ratings by analyzing a dataset containing variables such as budget, vote count, popularity, release date, and revenue. Then, by applying various statistical techniques, including linear regression, transformations, and interaction terms, to identify significant predictors and improve model performance. Throughout our analysis, we checked potential issues such as multicollinearity, violations of normality assumptions, and heteroscedasticity, while experimenting with variable selection methods like forward and backward selection, as well as Lasso regression.

Ultimately, our final model included vote count, budget, popularity, and release date as predictors, but it demonstrated limited explanatory power, with an adjusted R-squared of approximately 5.7%. This suggests that the selected predictors capture only a small fraction of the variance in movie ratings. Additionally, the model tended to overestimate ratings for recent movies, indicating potential biases in the dataset and limitations in its applicability to newer films.

For future research, incorporating more comprehensive and up-to-date data is essential. Variables such as social media engagement, audience reviews, and critic scores could provide additional insights into factors influencing movie success. Exploring non-linear models or machine learning approaches may also enhance predictive accuracy. While our study offers a foundational understanding of key predictors, it highlights the complexities of modeling subjective outcomes like movie ratings, underscoring the need for refined methodologies in future analyses.

Contribution:

- Multicollinearity & Model fit: Yutong Wu
- Variable Selection: Marie Picini Yufan Liu
- Interpret the Model:Yushan Guo