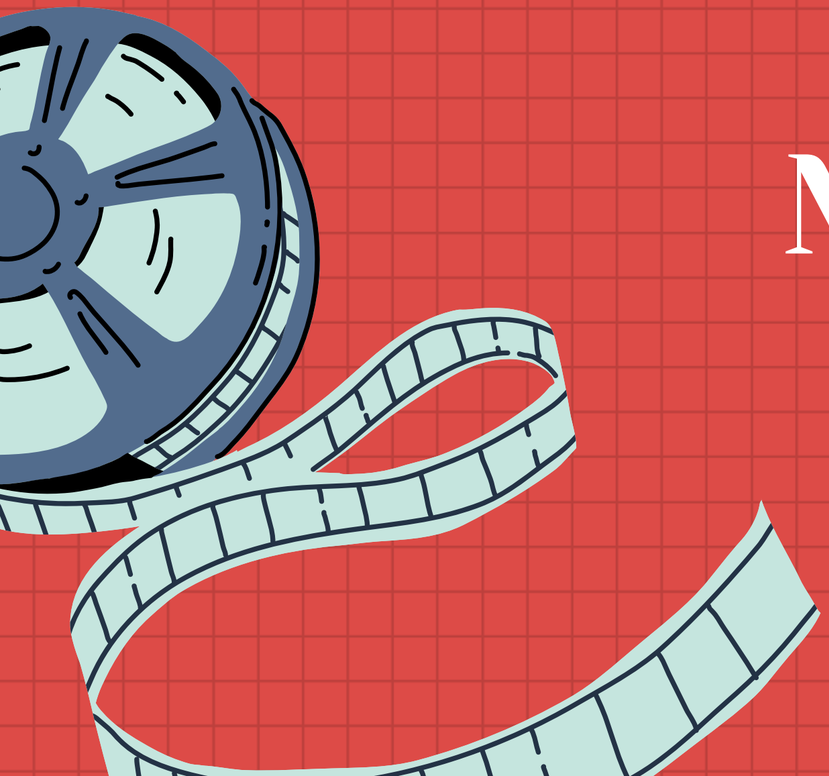
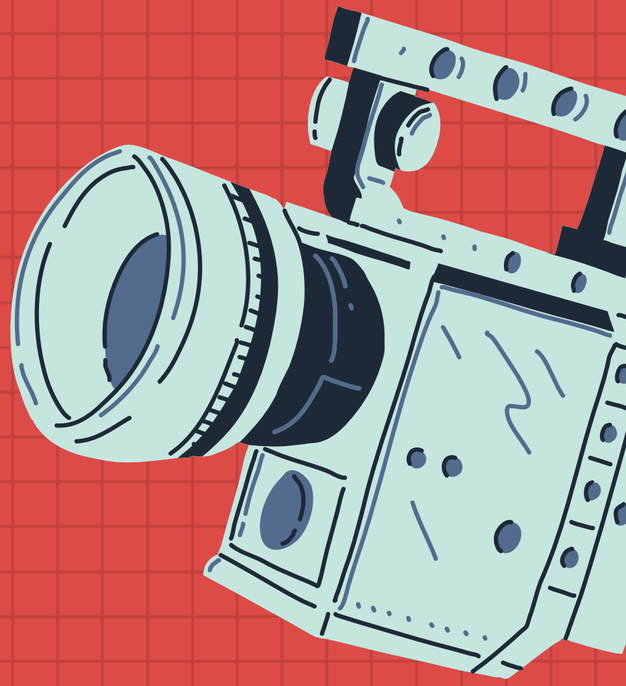
C3 Group1

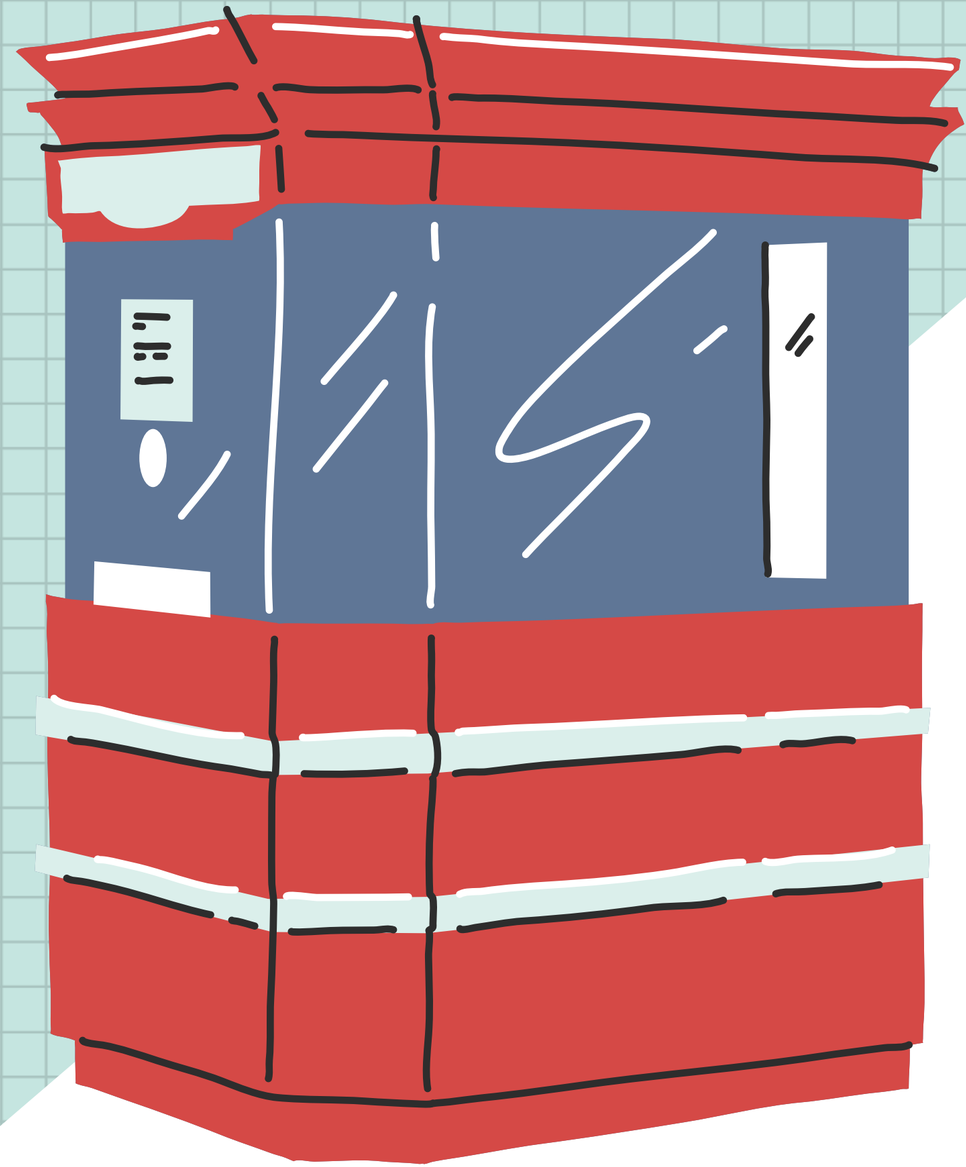
Primary Factors Influencing Movie Ratings

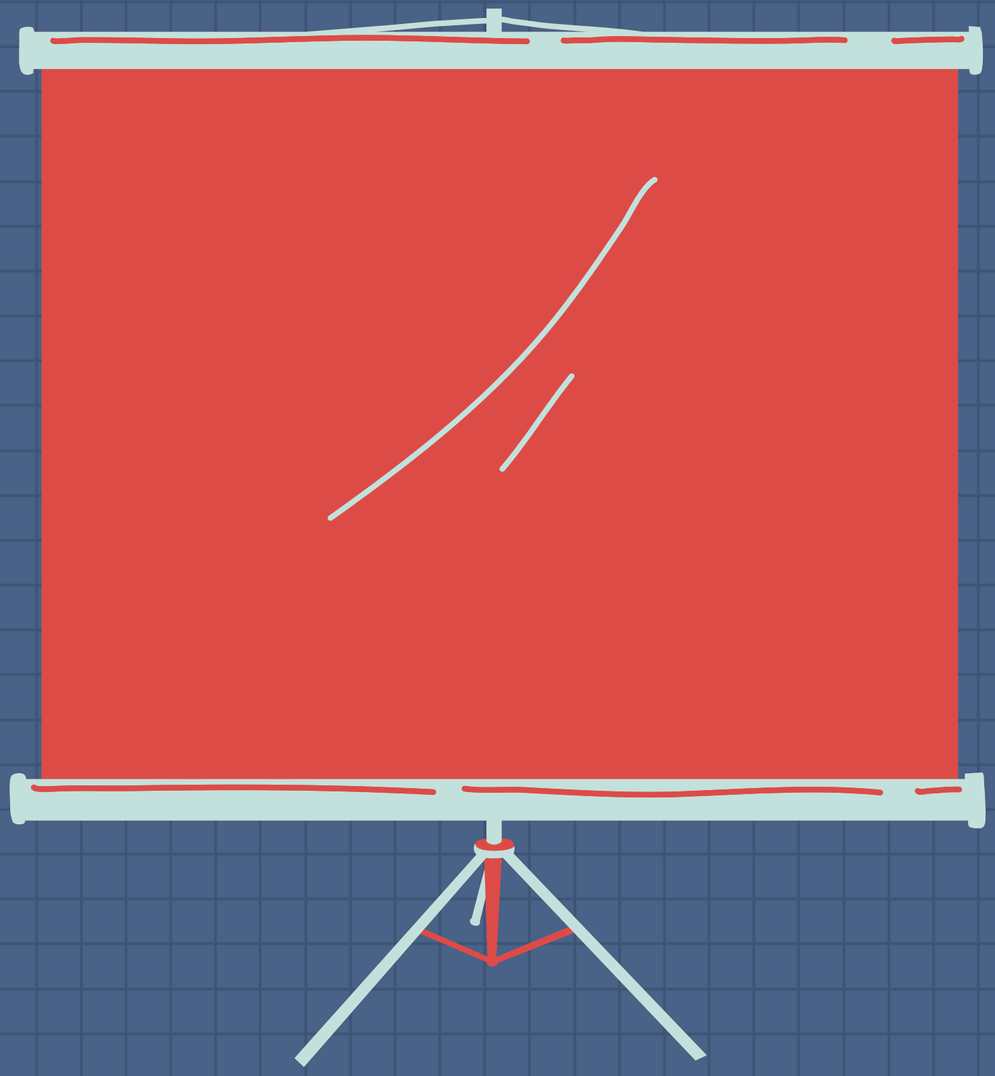
Yufan Liu, Marie Picini,
Yushan Guo, Yutong Wu





Index



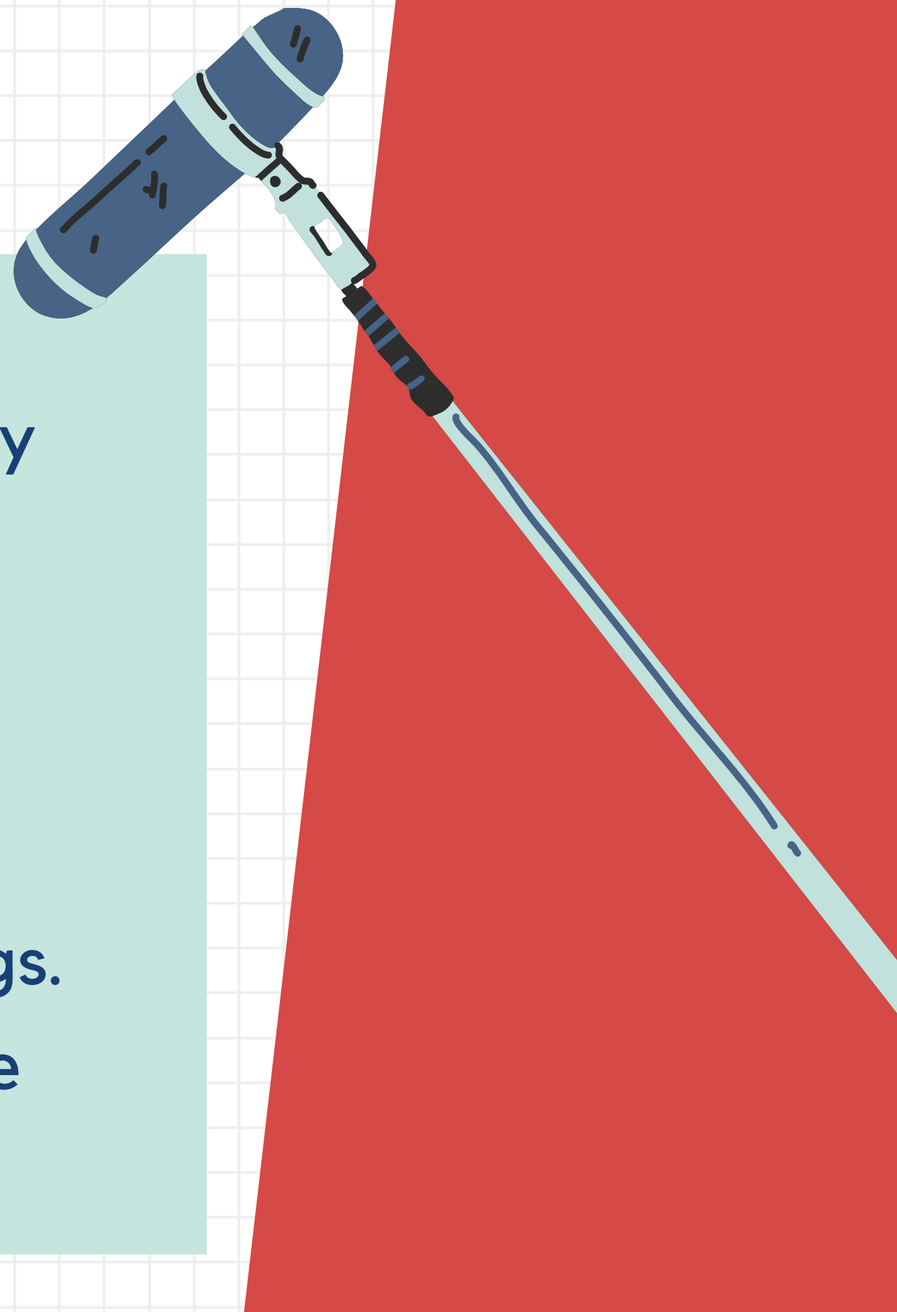


Introduction

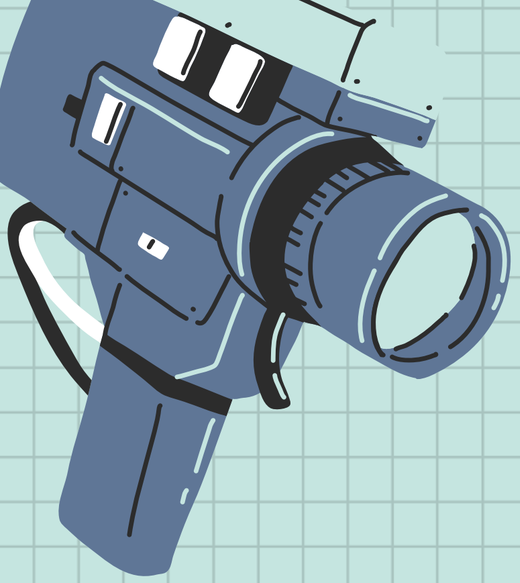
This project aims to identify key factors that influence movie success, focusing on the relationship between variables such as genre, budget, and director, and a movie's average rating (dependent variable).

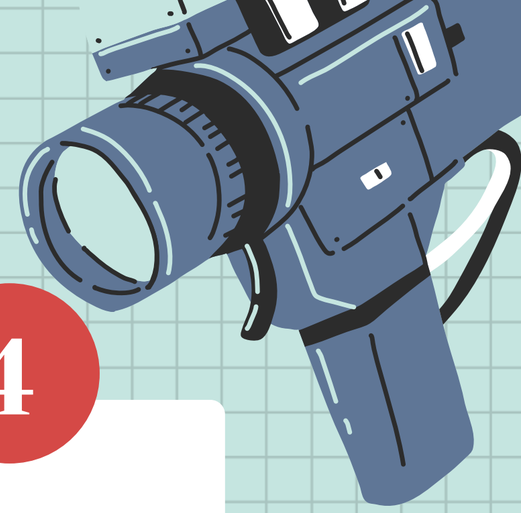
Hypotheses

- Higher revenue and budgets correlate with higher ratings due to better production quality and marketing.
- Action movies receive higher ratings due to their thrill-seeking appeal.
- Reputable production companies and renowned directors positively influence ratings.
- Cultural factors such as language and release timing impact audience reception.



Variables





1

Dependent (response) variable: vote_average (quantitative)

2

popularity (quantitative)

3

budget (quantitative)

4

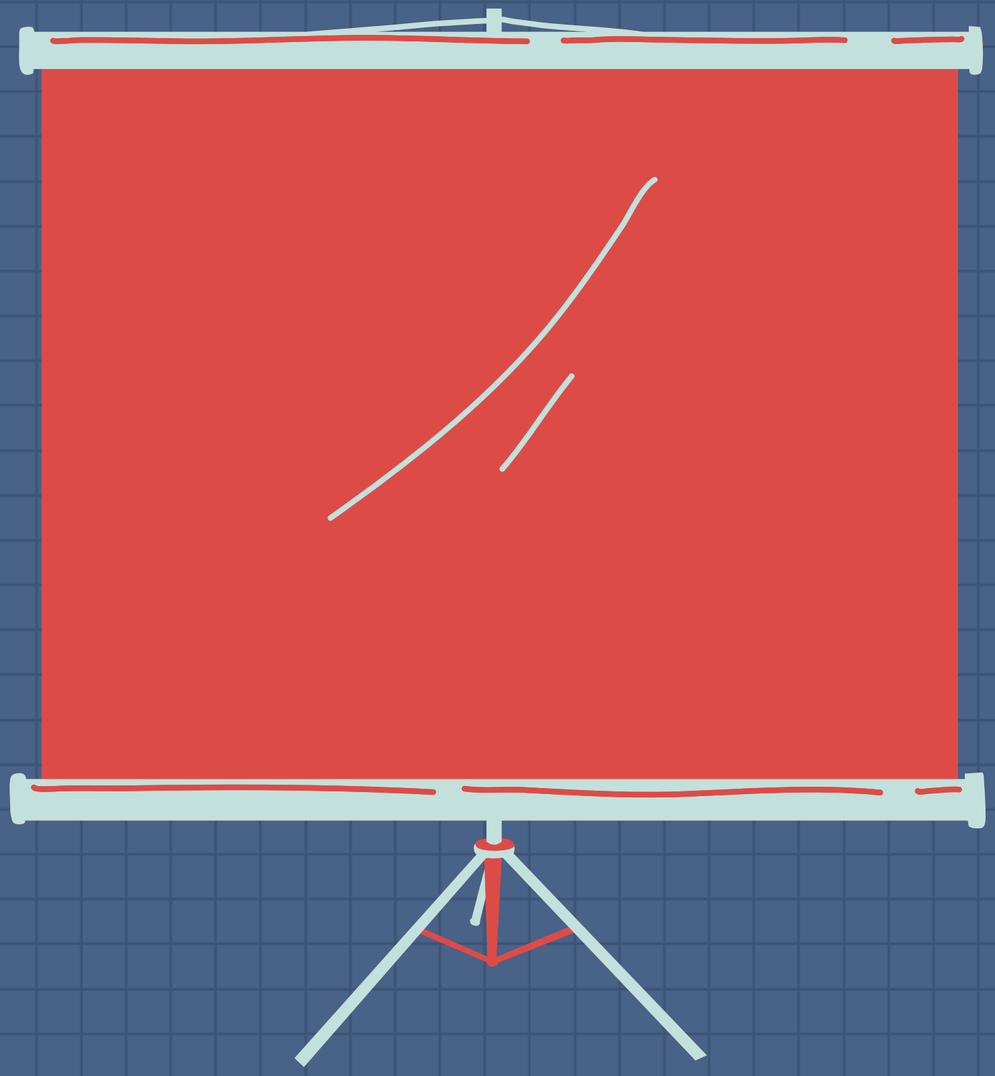
revenue (quantitative)

5

release_date (qualitative)

6

vote_count (quantitative)



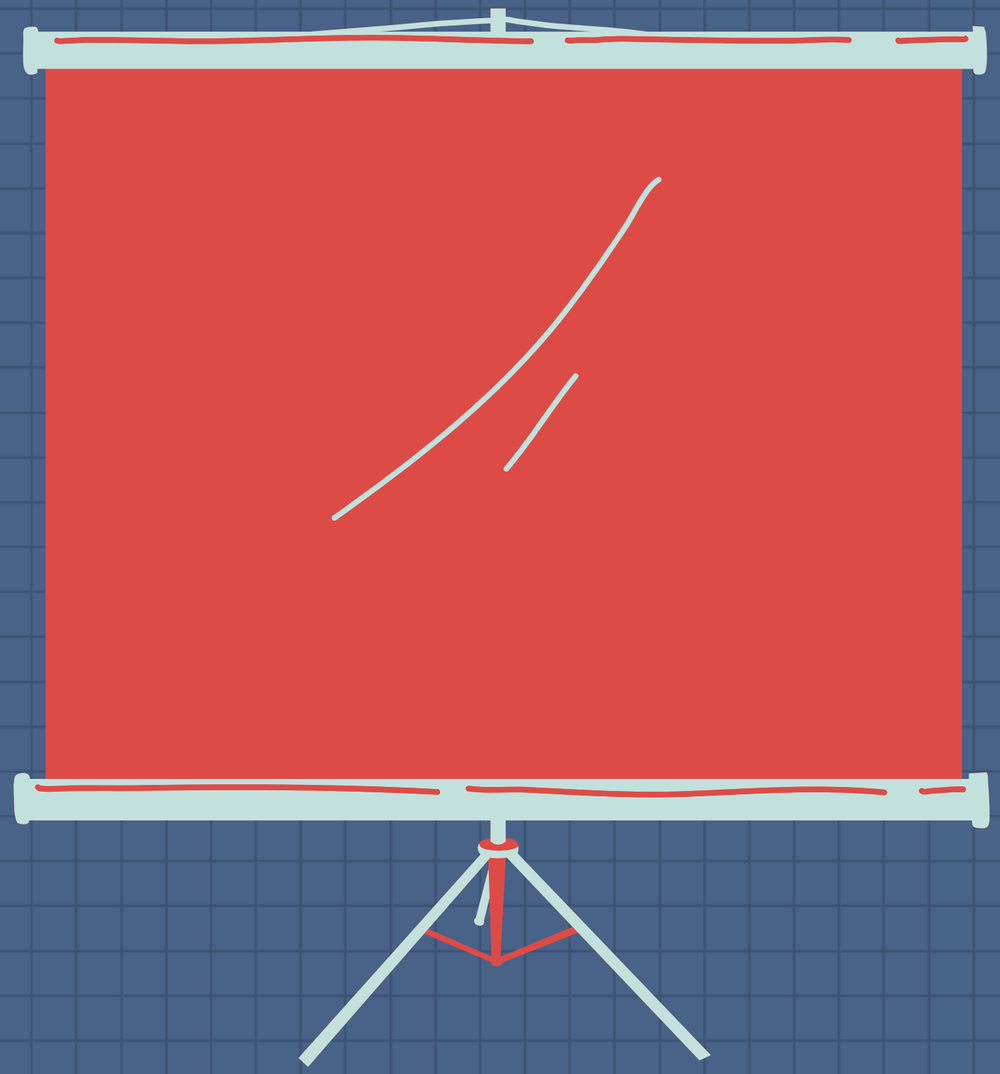
Full Model

Full Model

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

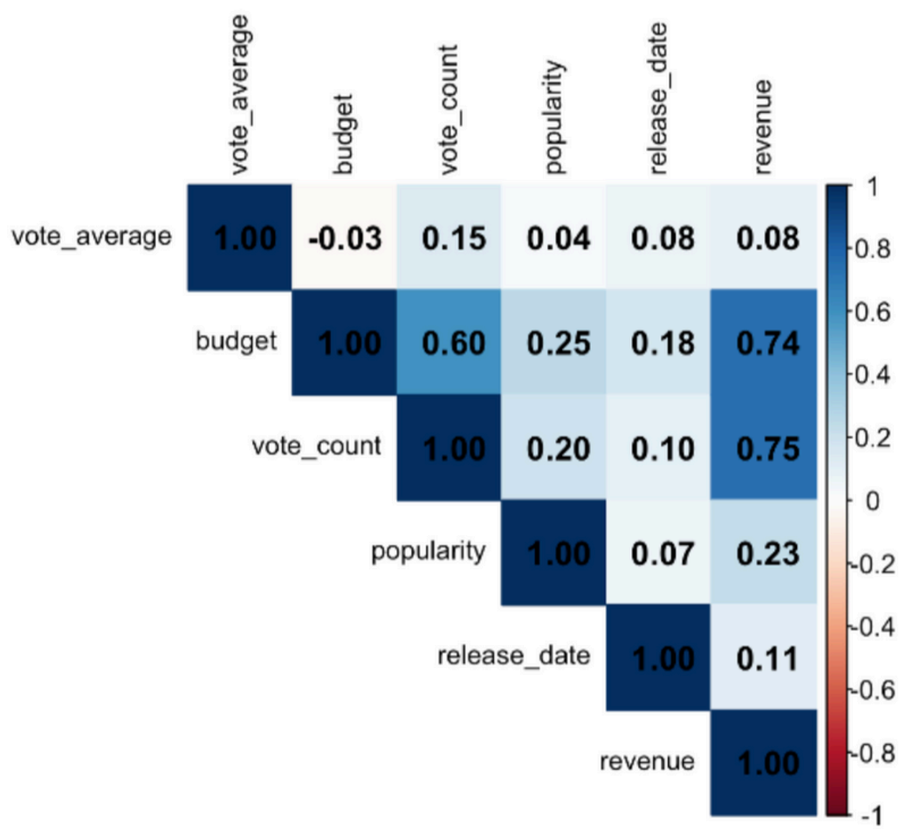
```
Call:
lm(formula = vote_average ~ budget + vote_count + popularity +
    release_date + revenue, data = cleaned_data)

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 R-squared -- only 5% of the variation in average vote scores is explained by the model.



Multicollinearity

Multicollinearity



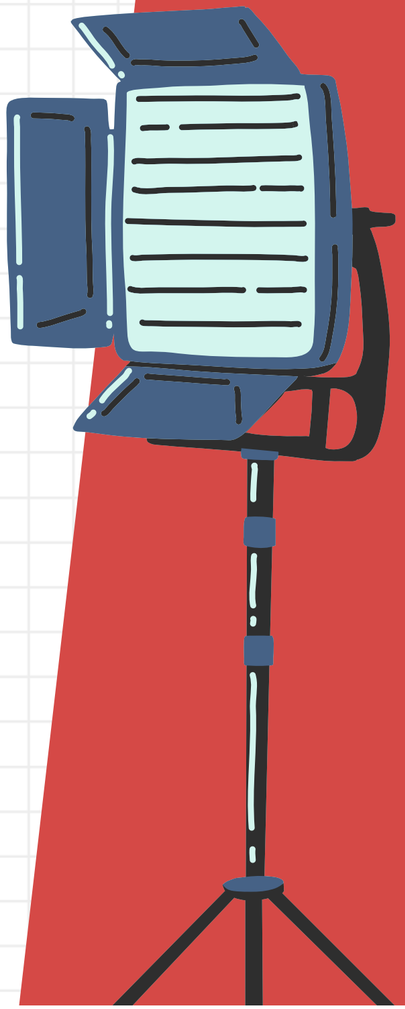
Budge, revenue, and vote_count all have two highly correlated covariates.

We tested variable removal and assessed its impact on adjusted R^2. Removing vote_count or budget lowered adjusted R^2, while removing revenue had minimal impact. Thus, we retained vote_count, popularity, release_date, and budget as covariates.

Model Fit

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 transformations below:

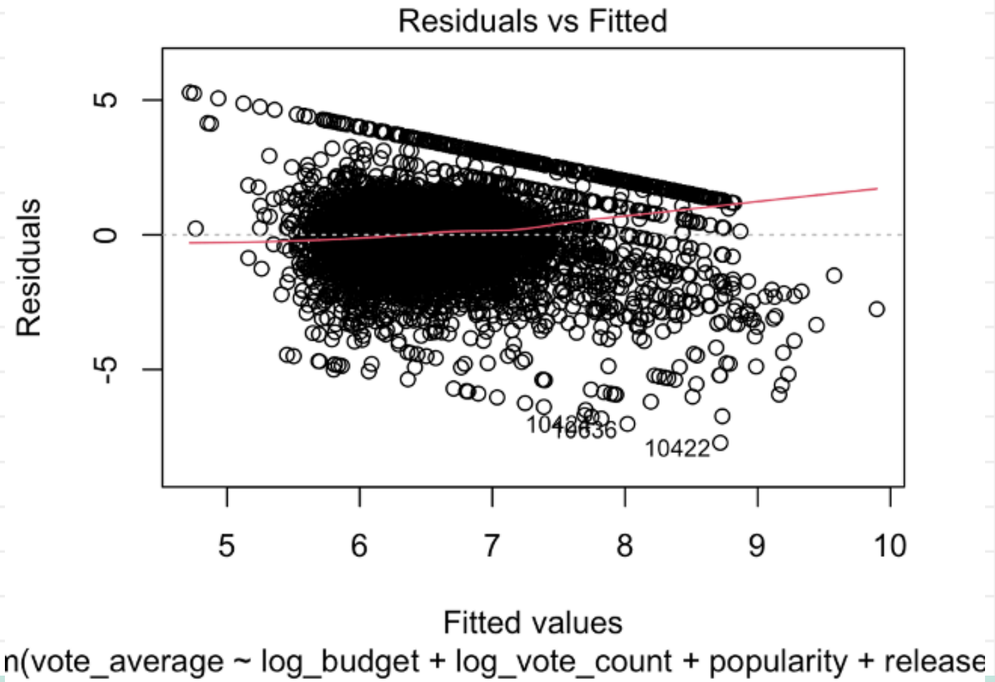
- Log-transforming skewed predictors budget and vote_count
- Including interaction terms: budget * popularity
- Cubic Polynomial Transformation
- Log and Cubic Polynomial Transformation



```
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
```

Log-transforming 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
```

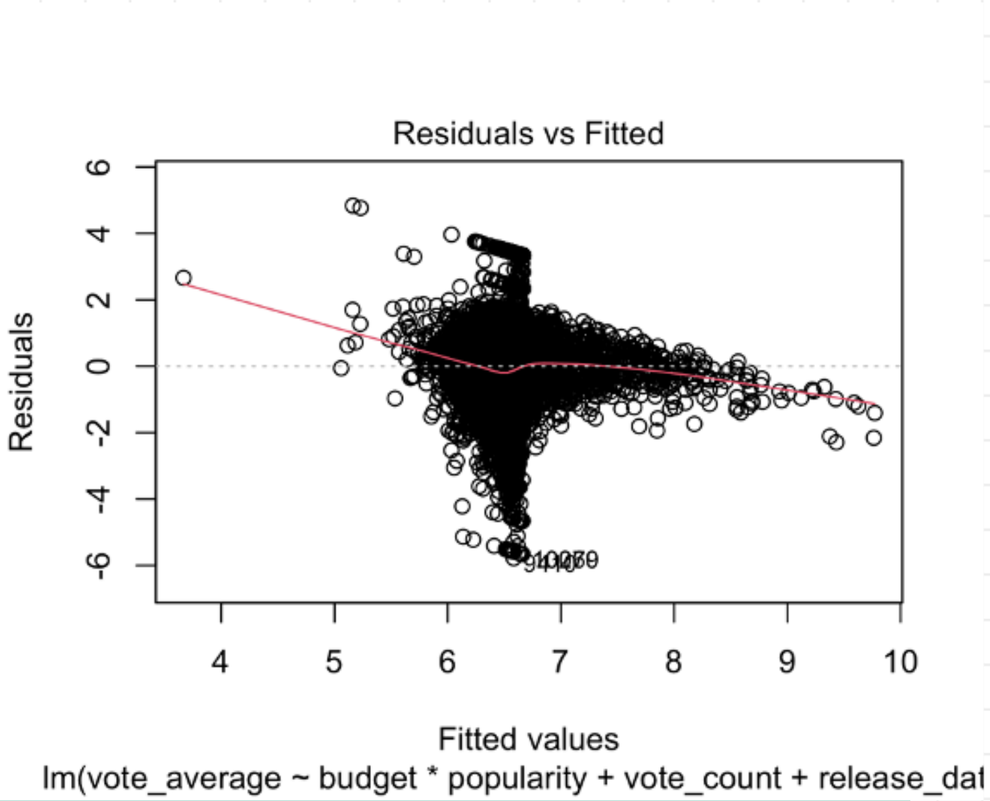


- **Improved adjusted R-squrared (0.1845 > 0.058)**
- **Residuals are centered around zero, but the plot suggests non-linearity and heteroscedasticity remain.**

Including interaction terms

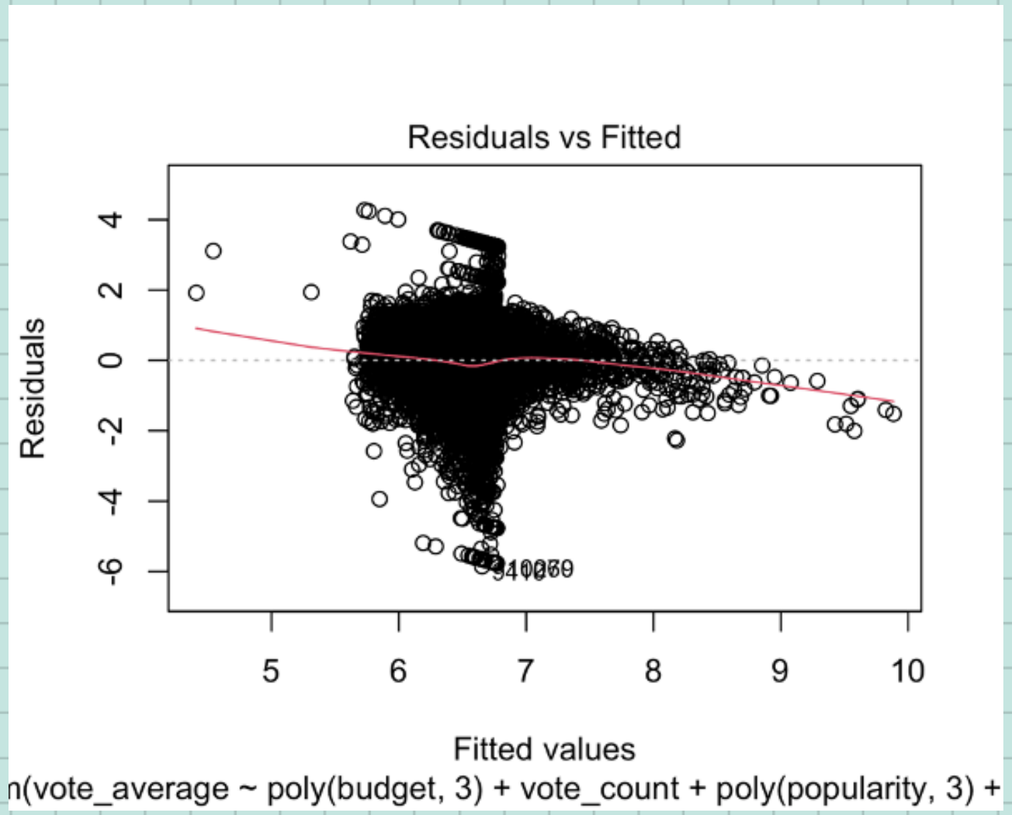
budget * popularity

```
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
```



- **Adjusted R-square is slightly higher than the baseline model(0.05939 > 0.058)**
- **Residuals showing signs of non-linearity and heteroscedasticity.**

Cubic Polynomial Transformation

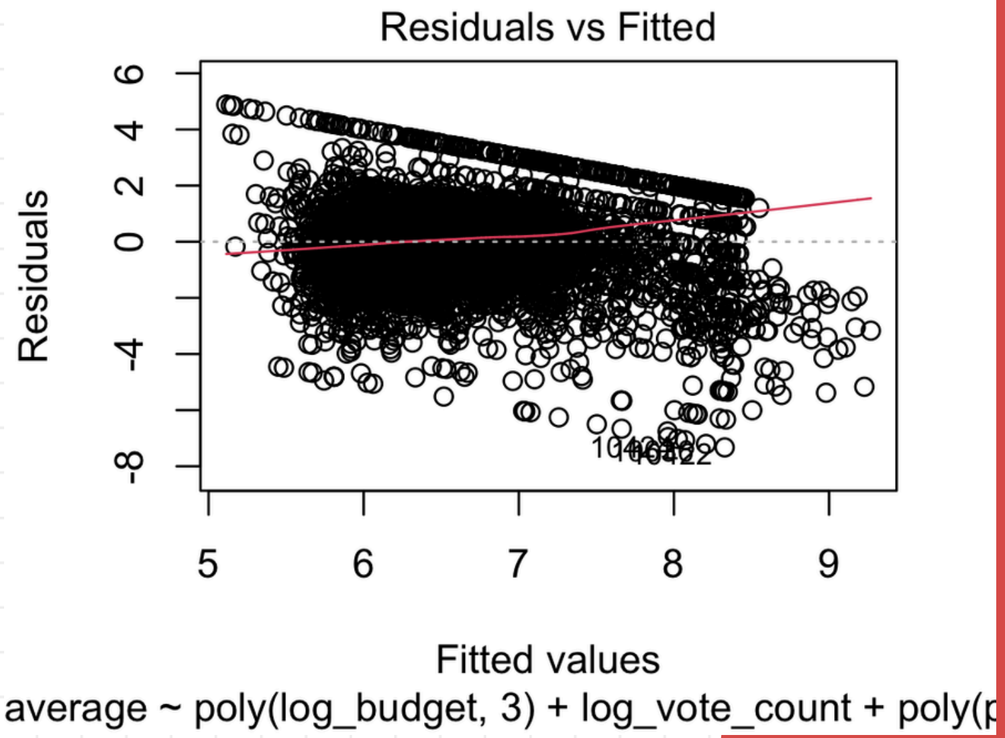



```
Call:
lm(formula = vote_average ~ poly(budget, 3) + vote_count + poly(popularity,
    3) + release_date, data = cleaned_data)

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
```

- **Adjusted R-square is slightly higher than the baseline model(0.07498 > 0.058)**
- **Residual plot shows improvement in capturing non-linear relationships, particularly for budget and popularity.**

Log and Cubic Polynomial Transformation

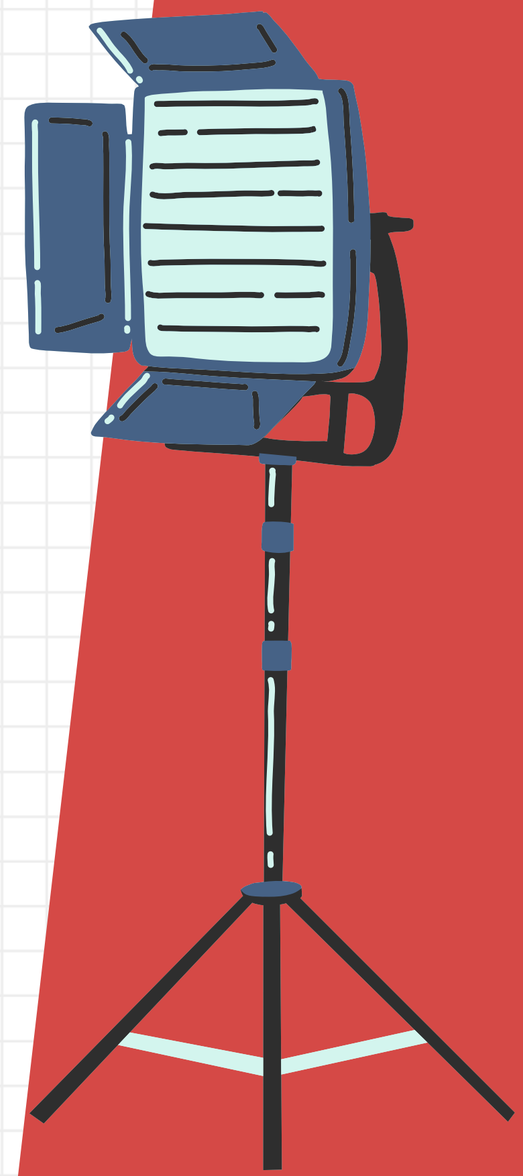
```
> summary(model_polylog)

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
```

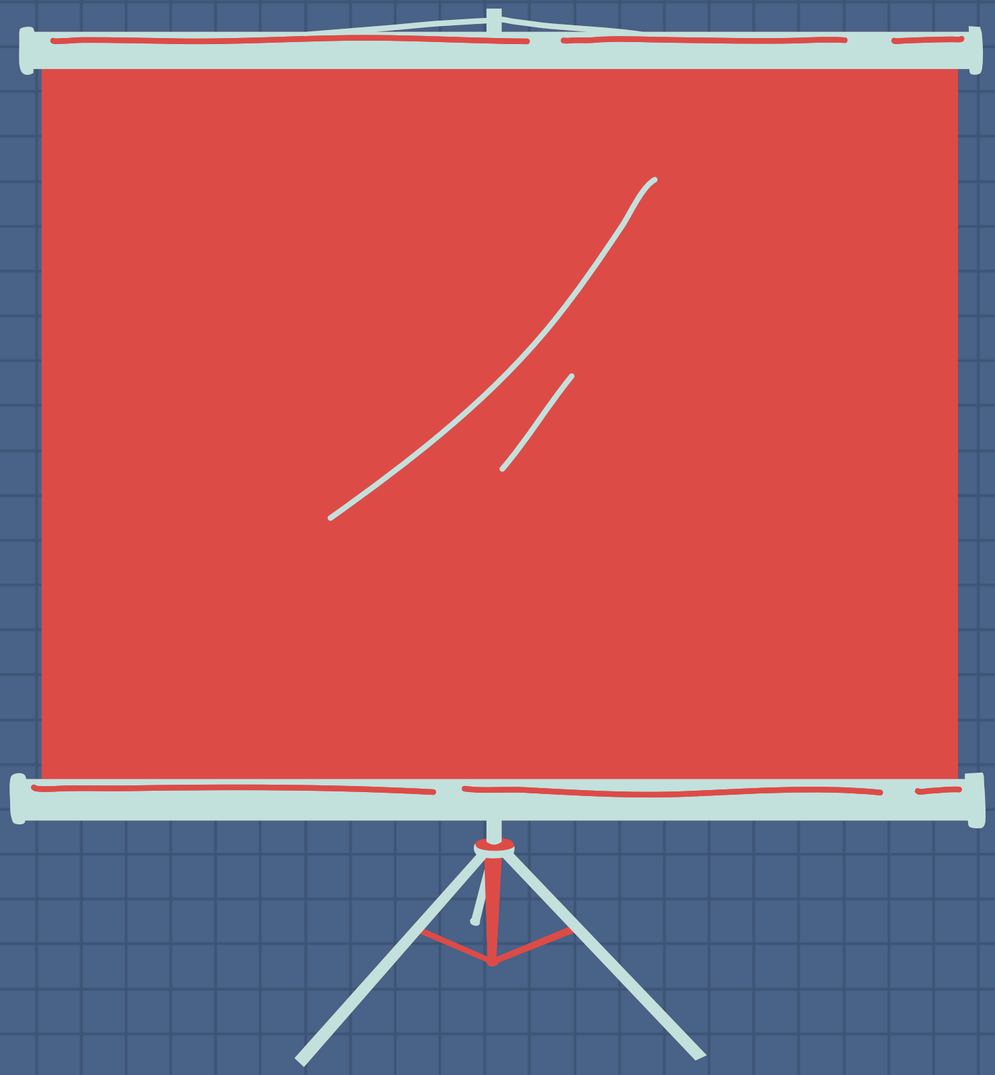


- **An improvement of adjusted R-squared over base model. (0.1998 > 0.058)**
- **Residuals showing signs of non-linearity and heteroscedasticity.**

Multicollinearity

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.





Variable selection

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

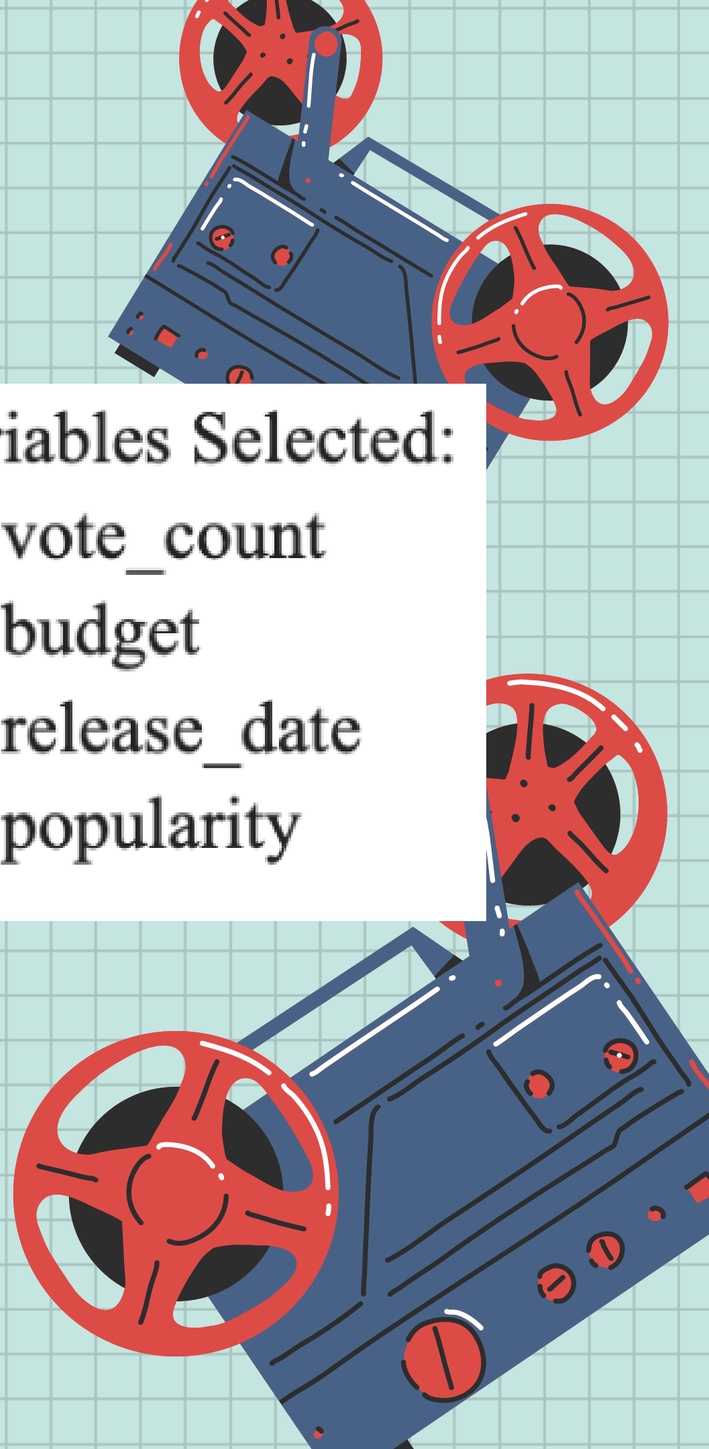
Variables Selected:

=> vote_count

=> budget

=> release_date

=> popularity



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.

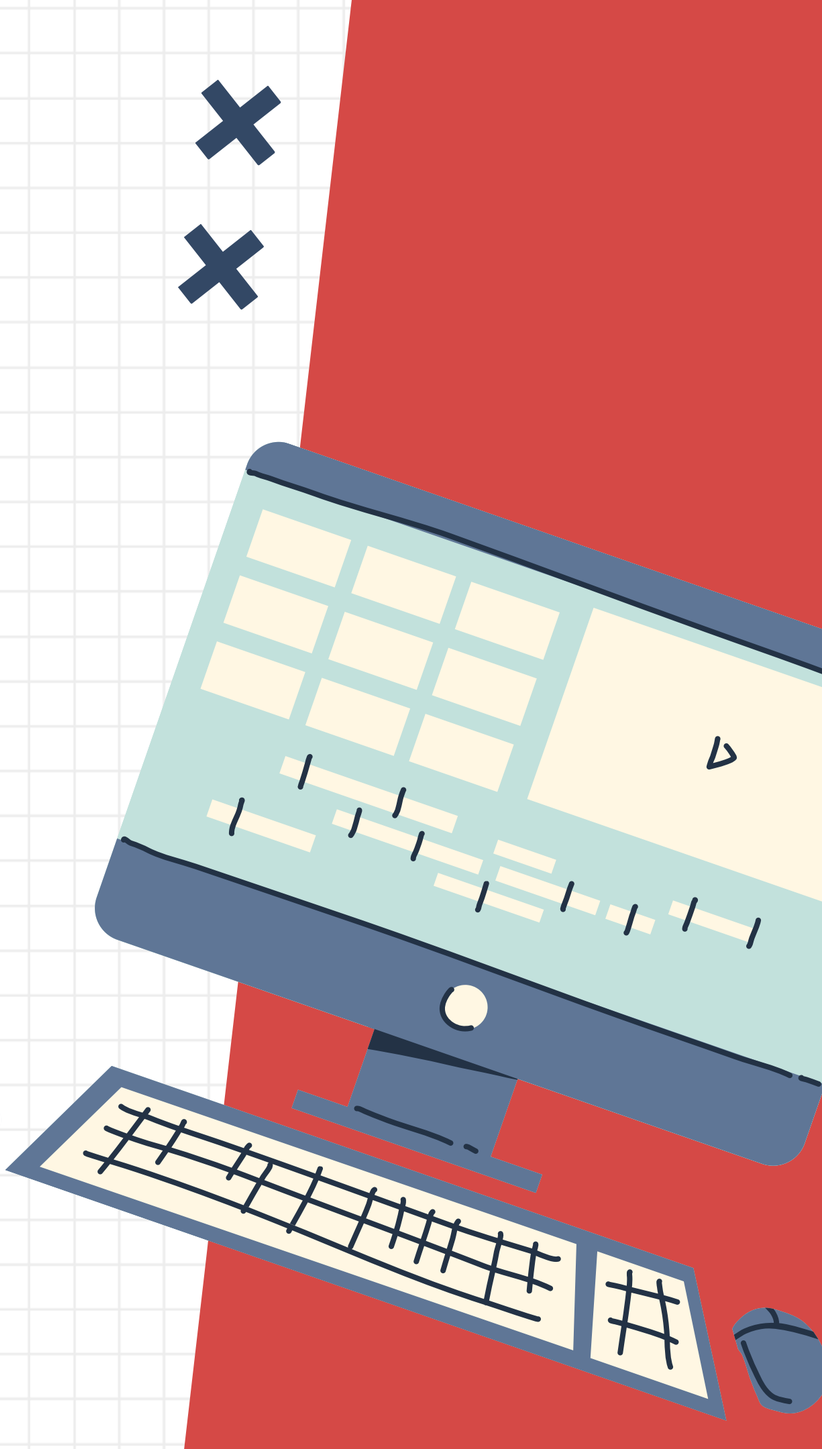
Because of our low results, I researched some other methods of variable selection and scaled regression:

LASSO

- Scales the coefficients of the model towards zero and eliminates informative variables in that process.
- The LASSO returned the same coefficients, meaning there wasn't much to any multicollinearity present within our variables and scaling wasn't really the main issue.

SCALE()

- scale() scales the data by the standard deviation, meaning the coefficients reflect the impact of 1 standard deviation increase in each predictor on the outcome variable.
- I used this method on independent variables then LASSOed again and found different coefficients.
- While they look more informative, in reality they are scaled differently, so they are just as uninformative and their effects are still negligible.



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

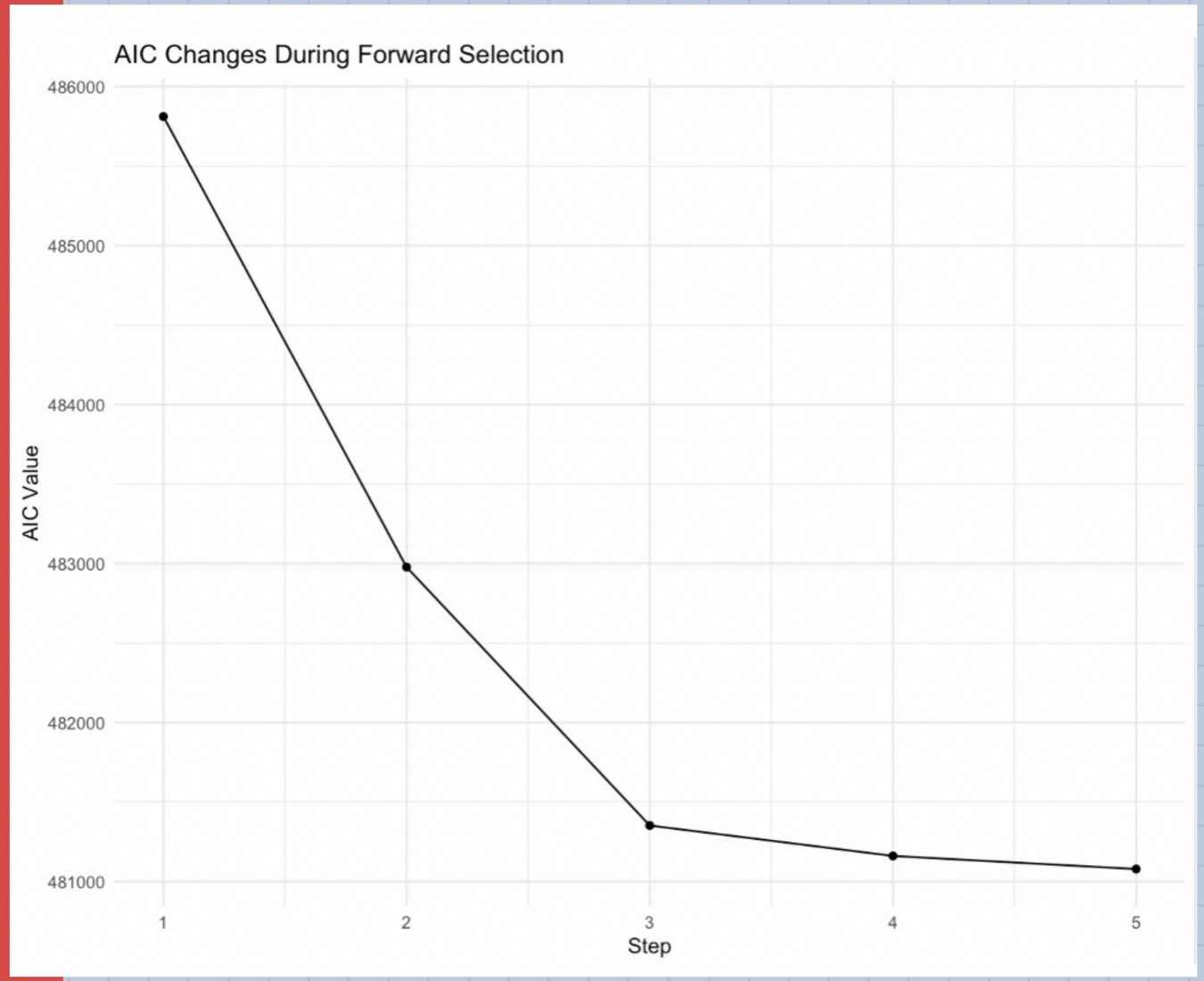
Scaled:

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

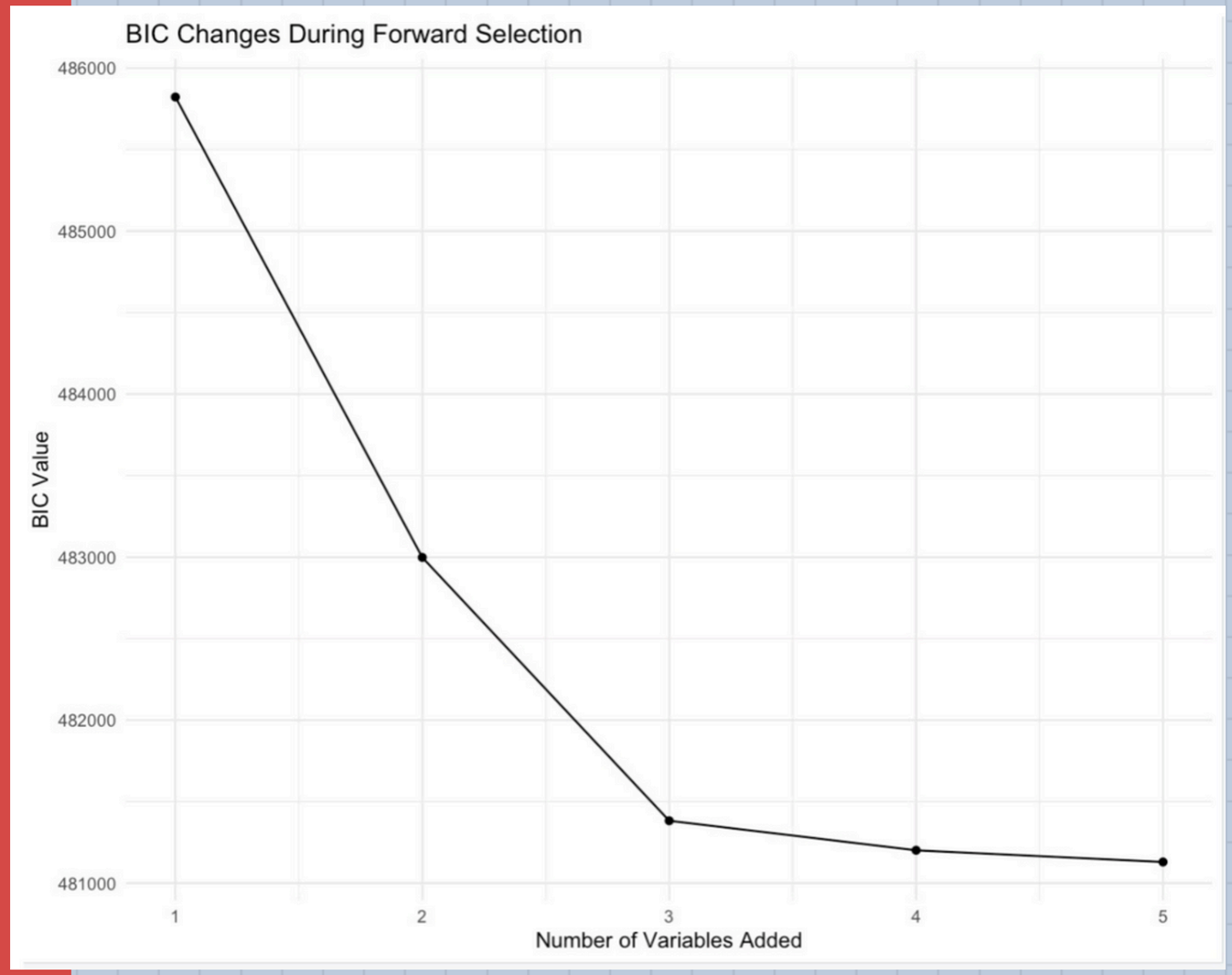
In addition to using the p-value criteria, for forward and backward, we tried AIC & BIC based criteria so that capturing and checking variables that should be included in the model.

Forward selection

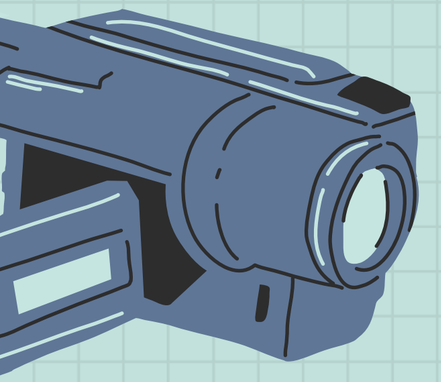


In the first few steps (steps 1 through 3), the AIC values drop quickly, indicating that these variables improve the model considerably.
From step 4 onwards, the AIC value decreases slower, indicating that the variables added later contribute less.

Forward selection

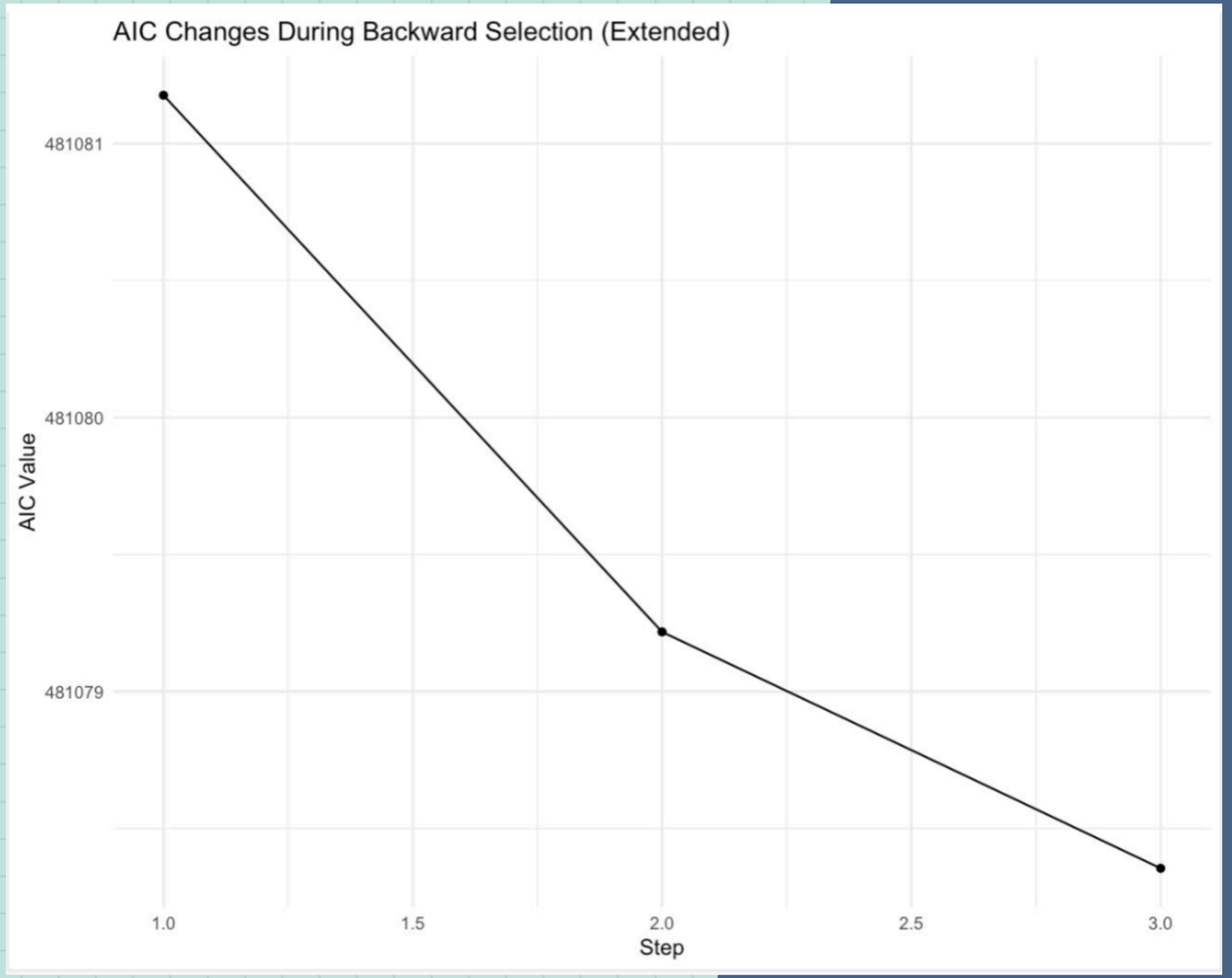


In the first few steps (from step 1 to step 3), the BIC value decreases rapidly, indicating that these variables significantly improve the model's performance.
Starting at step 4, the decrease in the BIC value becomes very small, suggesting that there is limited benefit from adding more variables.



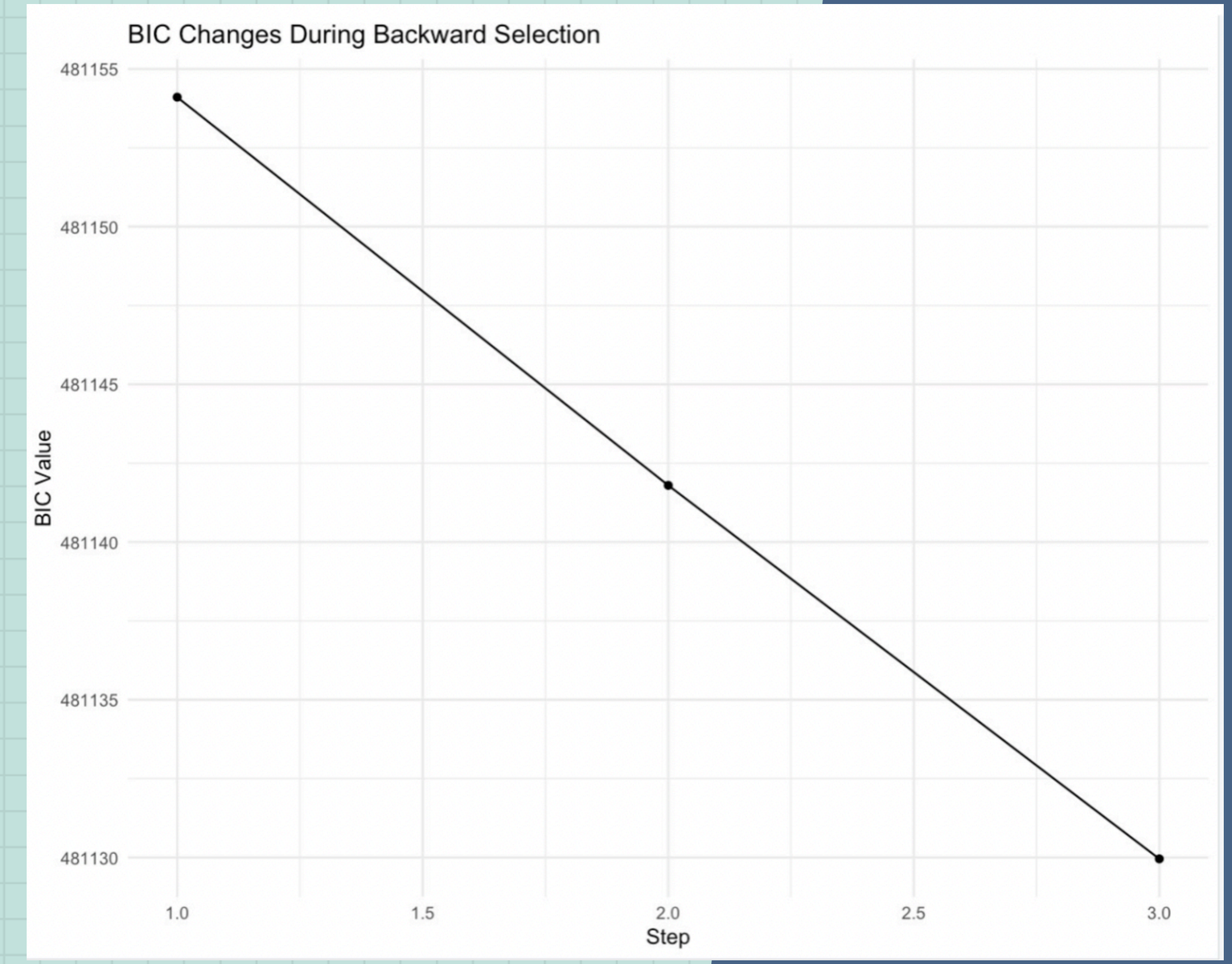
Backward selection

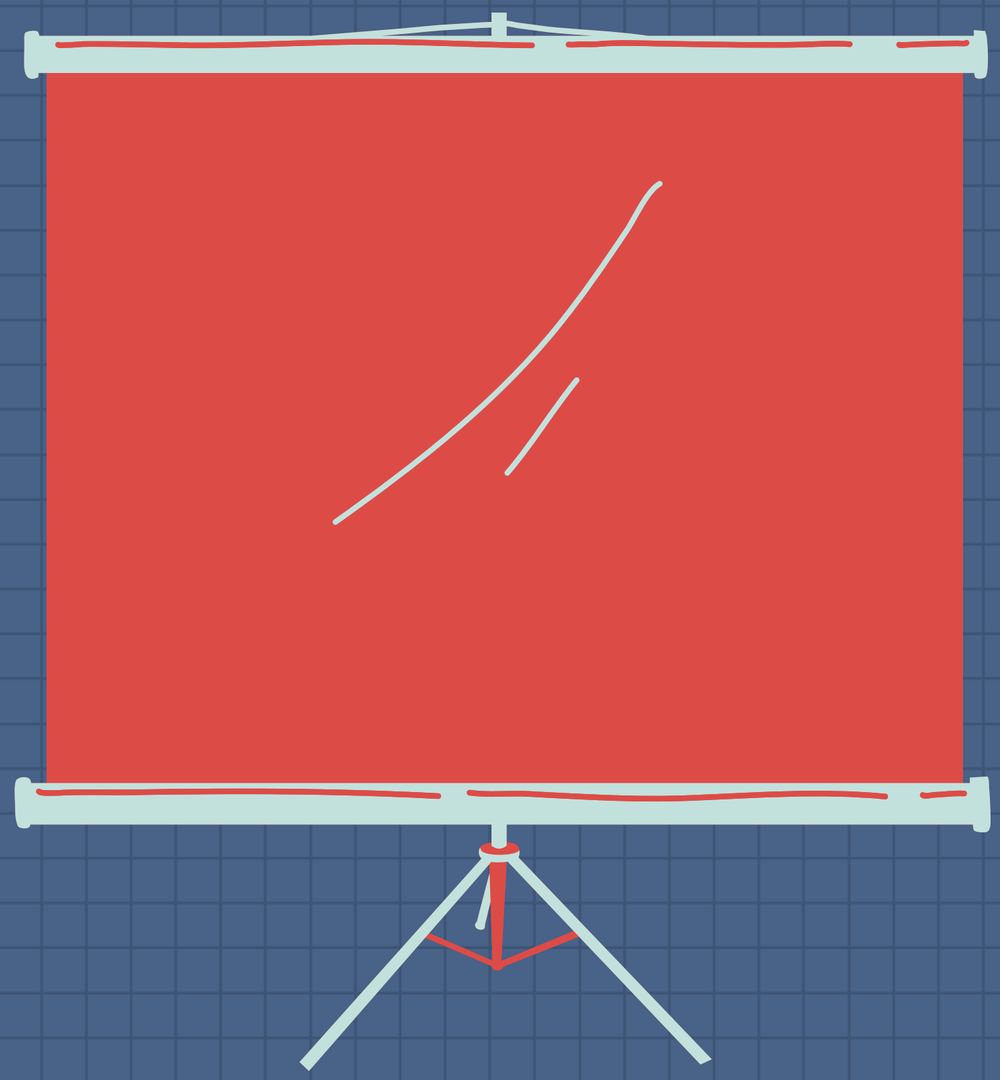
We can see that the AIC value gradually decreases from step 1 to step 3, indicating that the quality of the model is improving. The lowest AIC value is about 481,079, which occurs in step 3, which indicates that the model has reached an optimal equilibrium



Backward selection

We see a linear decrease in BIC values from step 1 to step 3, indicating that each removed variable has a similar effect on model simplification and quality improvement.
The lowest BIC value of approximately 481,130 occurs at step 3, indicating that the model is optimal at this point and achieves the goal of balancing complexity and explanatory power.





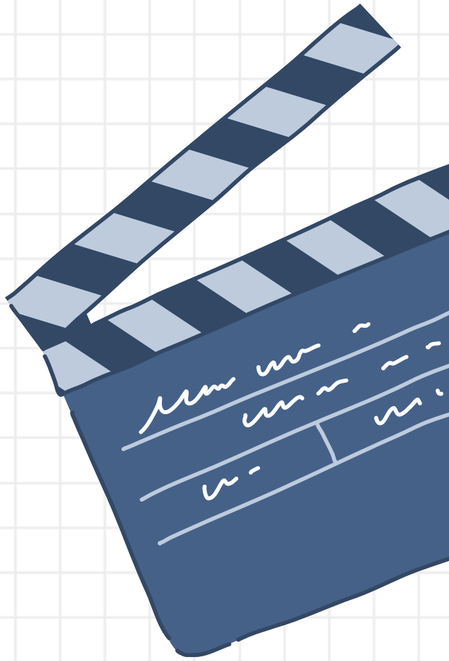
Model Interpretation

Natural interpretation?

Intercept: -5.401 --> **No**, because range for vote_average[0,10] & predictors like revenue, budget, released_date cannot be 0.

Slope --> Yes, but not very predictive of variations in vote_average
Budget: -0.000000007534
vote_count: 0.0001084
popularity:0.000743
release_date: 0.005805
revenue: 0.0000000004374

E.g For every one-year increase in the release year, vote_average increases by 0.005805, holding other variables constant.



What is predicting?

Consider a movie that has the following characteristics:

Budget: $10 million

Vote Count: 5,000

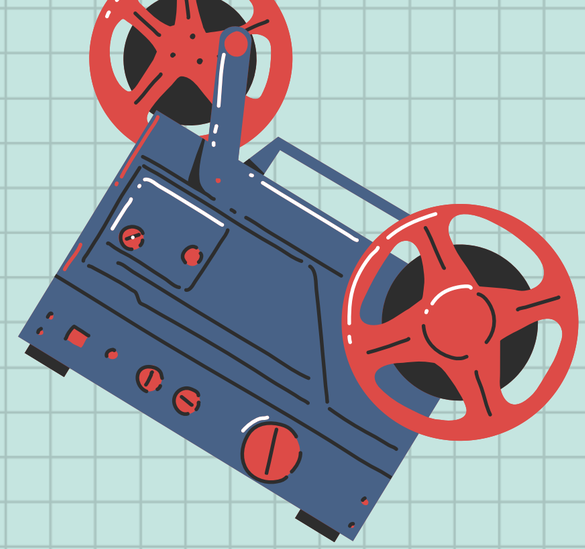
Popularity: 30

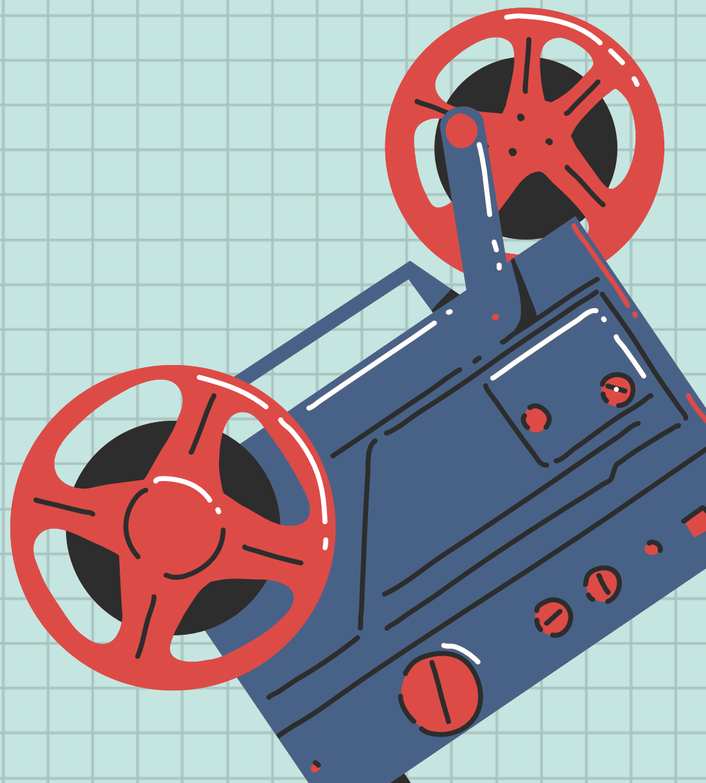
Release Date: 1985

Revenue: $50 million

vote_average = −5.097+(−0.000000007534×10,000,000)+
(0.0001084×5000)+(0.000743×30)+(0.005805×1985)+
(0.000000004374×50,000,000) = 6.937.

Also, all slopes are positive, meaning higher amount of predictors make vote_average to be higher.





Check Model Accuracy

There are two recent film that are not included in the dataset. We can use them to check:

Smile 2

Actual Rating from IMDb: 6.9/10
Expected Rating : 12.089(?)

Budget: $28 million, Release year:2024, Revenue:$130 million, Vote count: 50,759, Popularity: 119. (from IMDb)

95% C.I: (11.991, 12.187) -->not appropriate
95% prediction interval :(9.672, 14.506)-->not appropriate

Venom: The Last Dance

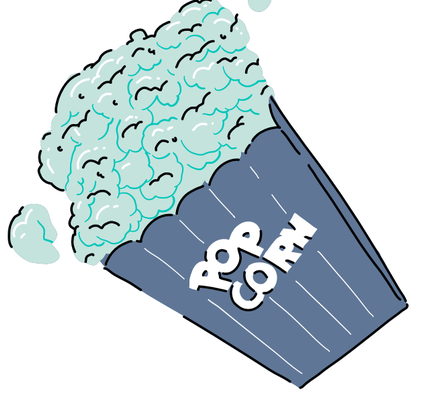
Actual rating from IMDb: 6.2/10
Expected Rating: 11.571(?)

Budget: $12 million, Release date: 2024, Revenue: $135,663,186, Vote count :45k, Popularity of 97(from IMDb).

95% C.I:(11.375, 11.767)
95% Prediction Interval : (9.148, 13.994).

Observation: Overestimation & Expected Rating exceeds 10.

Conclusion



Limited explanatory power

R^2 = 0.057, suggests 5.7% of variations in vote_average is explained by the model.

Limited Model accuracy

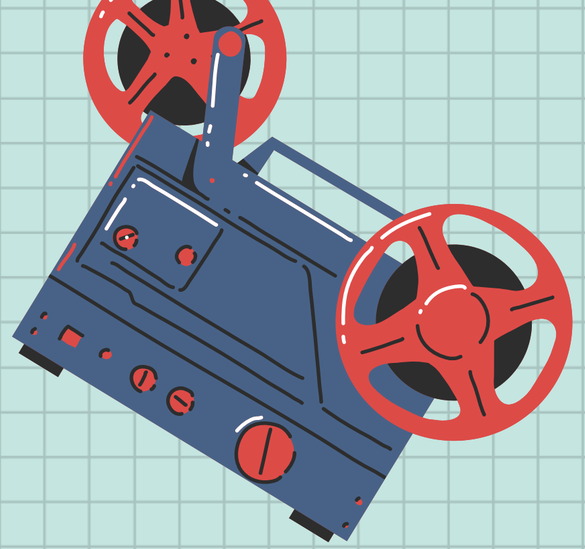
The predicted result deviates from the actual

Challenges

Little improvement in trying different variables & interaction term

Improvement

Trying different datasets
Incorporate variables such as social media engagement, audience reviews, and critic scores



Thanks for listening!

Questions?