What Makes a Movie Succeed?

Insights from Budget, Runtime, and Viewer Sentiment

Code Available @ https://github.com/arjashok/sta141b-final-project

Contributions:

Jack Ellis: Written portion of Abstract, Introduction, Methods, Results, Conclusion. Planning of API use/ideas for research questions.

Arjun Ashok: RQ1, RQ2 part ii; sentiment & review dataset generation; corresponding sections + editing

Jonathan Tran: Getting base dataset of movie ids, movie budget & rating/revenue dataset generation, RQ2 visualizations, T-tests, RQ2 ANOVA analysis

Matthew Chao: Conducted data cleaning, exploratory analysis, linear and quadratic regression modeling, and generated corresponding visualizations for RQ3

Prepared For:

Nicolai Amann

STA 141B

12 December, 2025

Abstract

This project aims to investigate whether key film attributes such as reviewer sentiment, production budget, and movie runtime relate to audience rating and financial performance using Python and The Movie Database's API. We collected data on all movies released in the United States between 2022-2024 with at least 500 user ratings. Starting with review comments, we used BERT for Natural Language Processing of the text in these reviews to determine that reviewer comments are generally more negative than the rating left by the reviewer. We performed t-tests to determine that higher budget movies, on average, succeed more in generating larger revenue and receiving higher ratings. A regression analysis showed that sentiment in review comments was not significantly different between high budget and low budget movies. Further regression analysis between movie runtime and the natural log of budget indicated that movie runtime is a positive predictor of movie budget. Another regression demonstrated that longer runtime is associated with higher ratings, although these regressions do not explain budget and ratings adequately, implying that other variables are better predictors.

Introduction

Throughout the last century and a half, people have fallen in love with movie watching as a hobby and social event. People have congregated all over the world at their nearest theaters to see stories come to life through motion pictures. Despite the overall success of movie watching as a whole, individual movie success can still be difficult to quantify, with a multitude of factors contributing to the final product. With movie budgets increasing higher than they ever have before in history, combined with a drop in theater visits due to the rise of streaming, we wanted to determine how the film industry is proceeding over the last two years. Our project aims to

determine whether the perceived audience rating of movies is significantly affected by the movie budget, longevity of the movie, and whether these number rating metrics are representative of the comments people feel about the movie. Accordingly, we ask the following research questions:

RQ1. Does the sentiment of reviews on imdb match the final review score? Concretely, can people translate their feelings of a movie into an accurate numeric score? Can we predict rating scores based on the sentiment of reviews?

RQ2. What is the average rating for films with a higher vs lower budget? What is the average revenue for films with a higher vs lower budget? How does budget relate to review sentiment/review score?

RQ3. How does the longevity of the movie affect the ratings? How does runtime relate to a movie's budget?

In order to answer these questions, we relied on Python as our scripting language to scrape, analyze, and display the data. We used The Movie Database's (TMDB) user-friendly API to collect all movie data, including movie budget, average rating, and user comments. We felt TMBD was an appropriate representation of movie ratings and budgets because it receives lots of user interaction and is considered one of the most popular movie rating websites. This approach also saved us time from parsing through the HTML of the website, giving us quick access to the important data. Once we retrieved all our data, we relied on Natural Language Processing (NLP) to convert text data into numerical data (via Transformers, PyTorch), which we then performed deeper analysis on. We displayed our data using the MatPlotLib and Seaborn libraries and used the StatsModels.API library to perform analysis such as regression and ANOVA. Utilizing these tools allowed us to present our significant findings via p-value and through visual plots.

Methods

Approach to Collecting Data

The data in this report was extracted from The Movie Database via their '/movie/' API endpoint. We decided to retrieve all movies from the years 2022 to 2024 by assigning the variables 'start date' and end date' to 2022-01-01' and '2024-12-31', respectively. We then created a function called 'get_movies' which took in these two dates as inputs and returned movie IDs filtered by 4 categories: movie language (English), region (US), release types (theatrical releases), and vote count (greater than 500). We only included movies with at least 500 ratings to remove small indie films with limited data. We called this API and performed all data analysis via Python, making sure to respect the API limits using `sleep(0.5)`.

RQ1: Review Sentiment VS Movie Ratings

We generated our list of reviews using the function `get_reviews`. This function iterates through the accumulated list of movie_ids and then sends requests through the `/movie/reviews` endpoint by first determining the number of reviews pages and then downloading each page, finally concatenating all reviews into a list of JSON objects. We then apply the 'transform_review' function to these reviews to extract all relevant identifiers before applying the 'accumulate_reviews' function to convert each individual review into a single dictionary of lists. These are all applied inside of the gen_review_dataset, which compiles the full dataset of reviews (`reviews.csv`) from the previously accumulated list of movie ids.

The next step was generating an unbiased scoring of the review's sentiment. Although a dictionary-based approach was appealing, the ability for deep learning methods to capture the complexities of natural language more accurately led us to select a BERT-based model from HuggingFace (nlptown's BERT-base multilingual review sentiment model). The model was trained to predict star ratings (1 to 5) from a review's text across 6 popular languages, including

English. Functionally, we treated this model as a pseudo-ground-truth for our research question given that its diverse training regime would optimize for an unbiased model, especially in comparison to an individual picking their rating. Since the TMDB's review scoring was (a) not always present and (b) gave ratings from 1 to 10, we did two key processing steps after running the model on all reviews:

1. We filtered out rows without a corresponding human rating score from TMDB
2. We mapped the human rating scores onto a 1 to 5 rating scale via $(X \,\%\, 2) + (X \,//\, 2)$, effectively pushing $\{1, 2\} \rightarrow$ 1 star, $\{3, 4\} \rightarrow$ 2 stars, etc.

Finally, we used the 'compare_review_scores' and 'prepare_comparison' functions to compare the predicted rating with the real rating to determine how accurate the reviews were at capturing the right sentiment. We also evaluate the alignment of our unbiased estimator (BERT) versus the human-generated scores through accuracy metrics and error distributions.

RQ2: Average Rating of Higher VS Lower Budget? Average Revenue of Higher vs Lower Budget? Budget Vs Sentiment Score?

To assess the question of whether higher budget movies on average perform better among audiences than low budget movies, we utilized a different API endpoint. Here, we were able to rely on the '/movie/details' endpoint, which we used in the 'fetch_and_categorize_movies' function to return a dataframe consisting of the following columns: id, title, revenue, rating, and budget category. Our budget category was defined as a cutoff at $50 million, with low budget movies being below this and high budget movies being $50 million or more. We decided on this cutoff by analyzing the budget consensus online, with StudioBinder defining a high-budget movie at $50 million or more. After retrieving the dataframe, we performed an ANOVA test, or t-test, between high and low budget movies on their audience rating and total revenue. To

determine whether the sentiment is different between high and low budget movies, we joined together the RQ1 sentiment analysis data frame using BERT and the budget data frame from 'fetch_and_categorize_movies.' After joining these datasets on their movie id, we performed a linear regression to see if the movie budget predicted a difference in sentiment.

RQ3: Longevity vs Movie Ratings? Longevity vs Budget?

To assess if movie longevity affects the audience ratings, we once again relied on the 'fetch_and_categorize_movies' function to return the movie runtime and its corresponding ratings. After obtaining these values, we performed EDA to determine whether we could apply a linear or quadratic model to the data. During the EDA, we cut out any unnecessary outliers that may interfere with the analysis using the 'remove_outliers' function, which ignored outliers exceeding the interval [Q1-1.5*IQR, Q3+1.5*IQR]. We analyzed scatterplots (Figure 3.1) and distribution plots (Figures 3.3 and 3.4) to better understand the data. The distribution plots revealed no extreme skew or irregular structure and supported our removal of outliers before modeling. The residual plot (Figure 3.6) shows random scatter with no discernible pattern, supporting the assumptions of linearity and constant variance. Now we perform linear regression and quadratic regression analysis.

In order to determine whether larger budget movies had a longer runtime, we once again relied on the dataframe generated from 'fetch_and_categorize_movies.' However, we joined this dataframe with the one from RQ2 so we could have all the features necessary to carry out our analysis. We again ran EDA (Figures 3.2 and 3.5) to determine whether or not a regression model would be appropriate, where we noted to apply a log transformation to our budget data. After taking the log of the movie budgets, we deemed a linear regression model appropriate for analyzing the relationship between budget and runtime.

Results

RQ1: Review Sentiment VS Movie Ratings

After BERT's analysis of the reviews and recoding of the scales to match, we found that the sentiment of the reviews was harsher than the ratings left on TMDB, suggesting that users are more forgiving in their ratings than their feelings imply. The correlation *(r=0.7)* was found to be sufficiently strong for prediction, but not perfect. Figure 1.1 demonstrates the strong correlation between the sentiment and ratings, with the color coding demonstrating the model's confidence. Additionally, a large majority of the reviews were within one star of the rating assigned on TMDB, indicating BERT was relatively accurate in its assignment of stars (Figure 1.2).

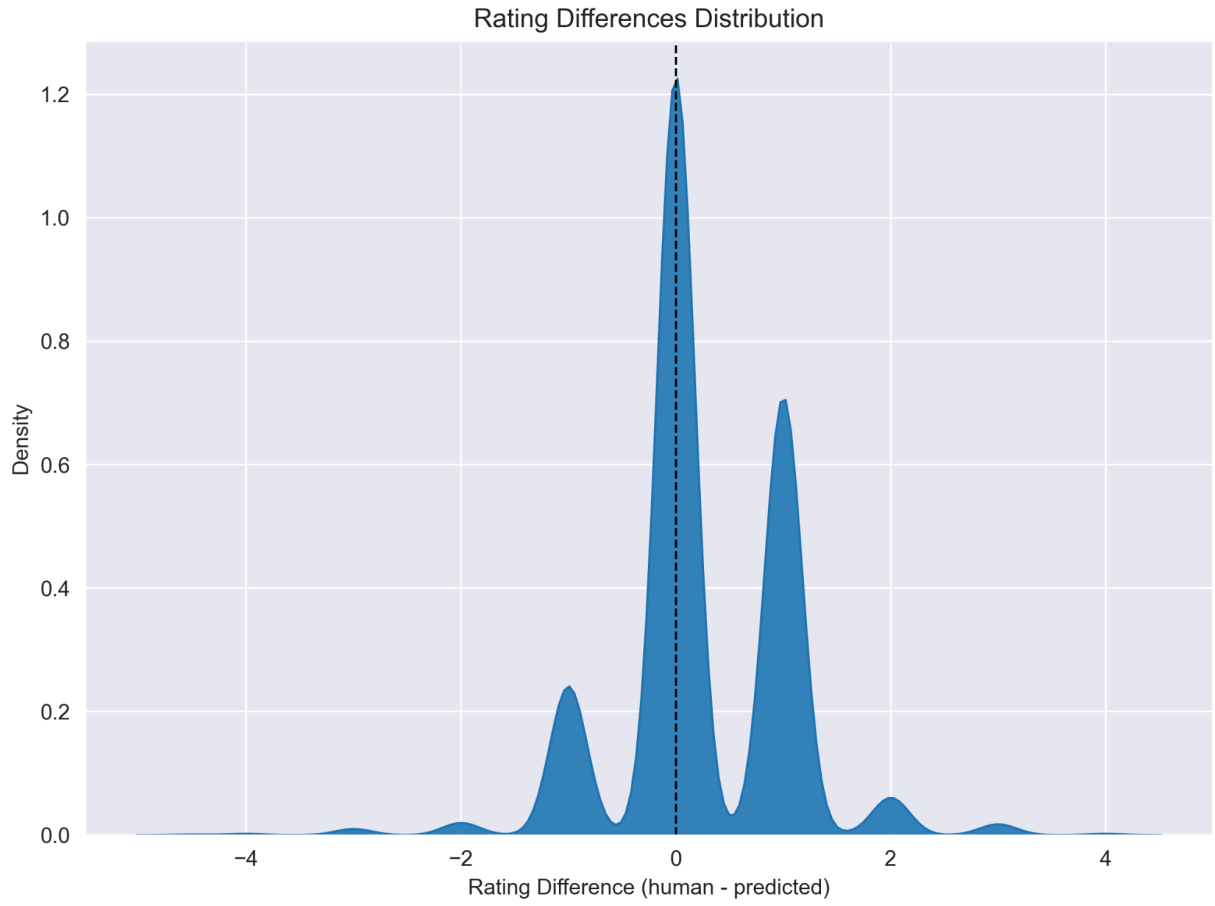


Figure 1.2: Difference Between BERT Rating and Human Rating

RQ2: Average Rating of Higher VS Lower Budget? Average Revenue of Higher vs Lower Budget? Budget Vs Sentiment Score?

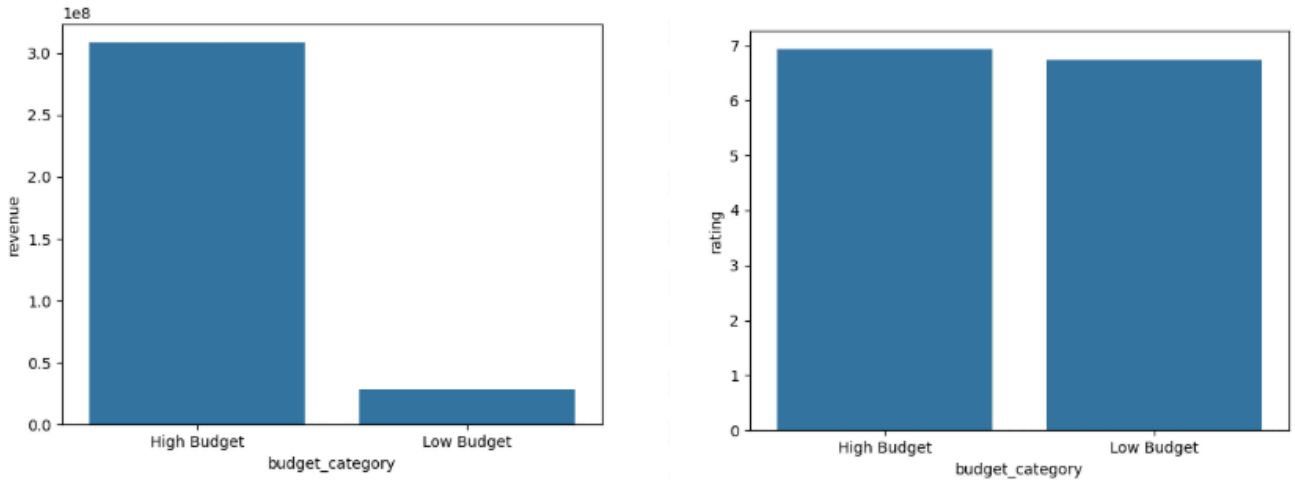


Figure 2.1: Bar Chart Comparison of Higher vs Low Budget Films' Revenue/Rating

We obtained a mean rating of 6.93 for high budget movies and 6.73 for low budget movies. We then ran a t-test to determine whether this difference in scores was significant, where we observed a statistically significant difference ($F(1,439) = 10.68$, $p = .001$). Since our p-value is lower than .05, we conclude with 95% confidence that there is a significant difference in ratings between high budget and low budget movies, with higher budget movies yielding higher ratings.

	sum_sq	df	F	PR(>F)
C(budget_category)	3.697967	1.0	10.676296	0.00117
Residual	152.057188	439.0	NaN	NaN

Figure 2.2: ANOVA Table For Budget vs Rating

We obtained a mean revenue of 308 million for high budget movies and 28.6 million for low budget movies. We then ran a t-test to determine whether this difference in revenues was significant, where we observed a statistically significant difference ($F(1,439) = 151.33$, $p <$

.0001). Since our p-value is lower than .05, we conclude with 95% confidence that there is a significant difference in revenues between high budget and low budget movies, with higher budget movies yielding higher revenues.

	sum_sq	df	F	PR(>F)
C(budget_category)	7.539566e+18	1.0	151.329994	4.353880e-30
Residual	2.187187e+19	439.0	NaN	NaN

Figure 2.3: ANOVA Table For Budget vs Revenue

We obtained a mean sentiment of 3.23 for high budget movies and 3.15 for low budget movies. We then performed a linear regression analysis to determine whether budget was a significant predictor of sentiment, but we failed to find a correlation between the two variables. Interestingly, Figure 2.4 implies that higher budgets create more consistent, less risky volatile ratings on average; observe the cone-like shape as the budget increases.

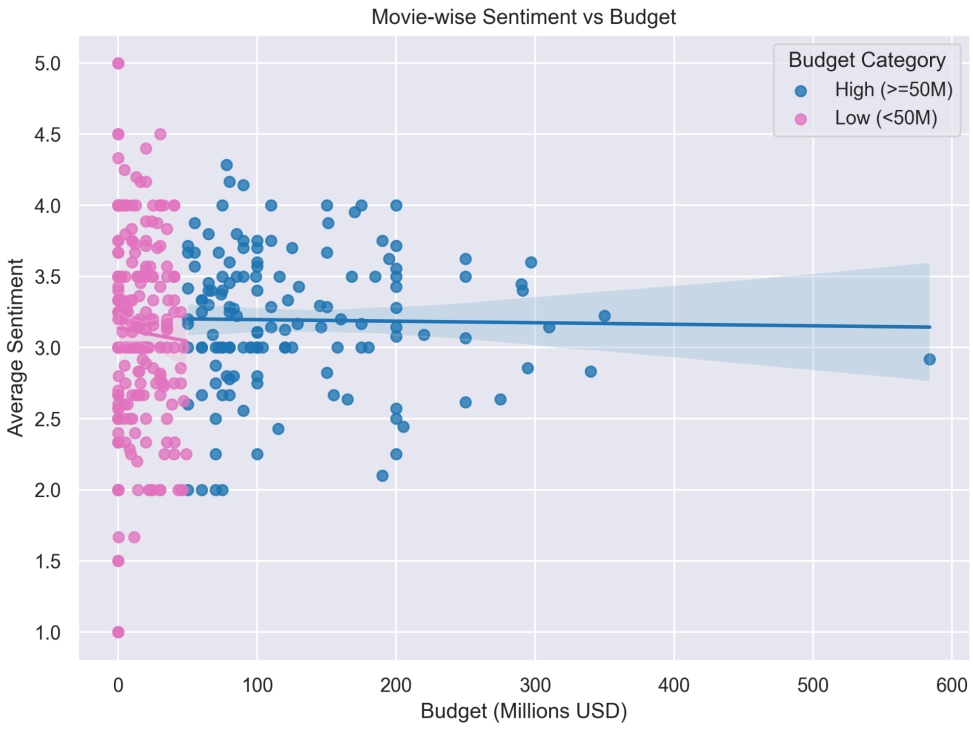
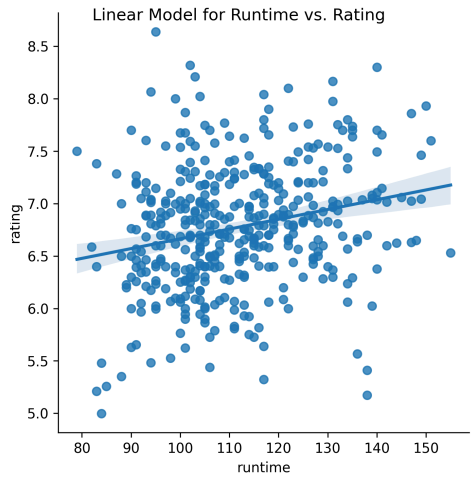


Figure 2.4: Linear Regression On Budget vs Sentiment

RQ3: Longevity vs Movie Ratings? Longevity vs Budget?

After fitting a linear regression model, runtime was found to be significant (t(414) = 4.98, $p < .0001$). This indicates that runtime does have a significant effect on a movies' rating. On average, for every additional minute of runtime, the predicted movie rating increase is about 0.0093. However, an R^2 of 0.06 means that runtime only accounts for 6% of the variation in ratings. (Figure 3.7) This means runtime is a weak predictor of movie rating. Likely other factors, such as budget or cast, for example, may account for the majority of rating variation.

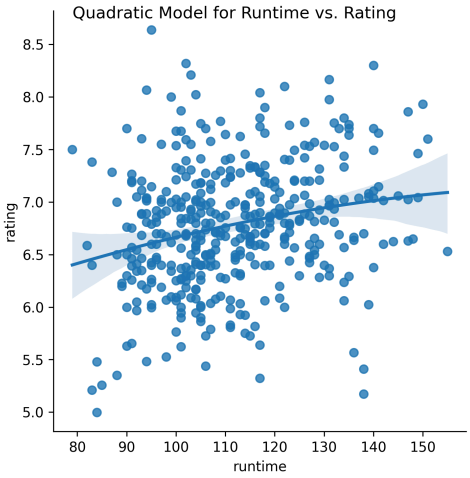


Dep. Variable:	rating	R-squared:	0.056
Model:	OLS	Adj. R-squared:	0.054
Method:	Least Squares	F-statistic:	24.77
Date:	Sun, 07 Dec 2025	Prob (F-statistic):	9.47e-07
Time:	00:43:46	Log-Likelihood:	-358.42
No. Observations:	416	AIC:	720.8
Df Residuals:	414	BIC:	728.9
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	5.7311	0.211	27.175	0.000	5.317	6.146
runtime	0.0093	0.002	4.977	0.000	0.006	0.013

Figure 3.7: Runtime v. Rating Linear Model and Results

To see if we can improve the R^2, we decided to run a quadratic regression model. After fitting the model, runtime and runtime squared were found to be insignificant (t(413) = 0.02, p = .32), (t(413) < 0.000, p = .55). This disagrees with the linear regression finding that runtime is a significant predictor of movie rating, demonstrating that a more complex model does not adequately explain this relationship. The R^2 of .06 remained the same as the linear regression model, further illustrating that a more complex model is not explanatory of the difference in ratings. (Figure 3.8) Therefore, other variables should be considered to determine what predicts ratings left on movies.

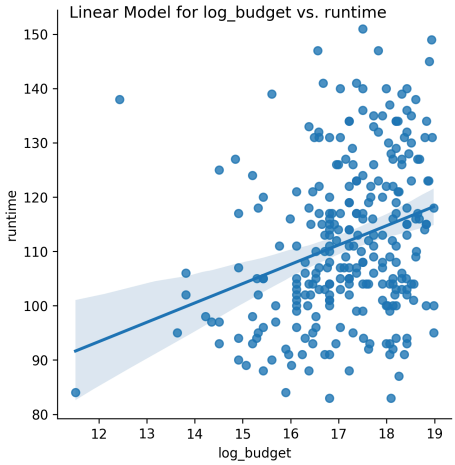


Dep. Variable:	rating	R-squared:	0.057
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	12.55
Date:	Sun, 07 Dec 2025	Prob (F-statistic):	5.11e-06
Time:	00:43:47	Log-Likelihood:	-358.24
No. Observations:	416	AIC:	722.5
Df Residuals:	413	BIC:	734.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.9084	1.376	3.567	0.000	2.204	7.613
runtime	0.0239	0.024	0.989	0.323	-0.024	0.071
runtime_sq	-6.332e-05	0.000	-0.605	0.545	-0.000	0.000

Figure 3.8: Runtime v. Rating Quadratic Model and Results

Since we determined movie longevity was not adequate in explaining the differences in ratings, we were interested in determining whether longevity could explain the differences in movie budget. To assess this, we again ran a linear regression on the log budget vs movie runtime. This regression indicates that runtime does have a significant effect on the log of a movie's budget ($t(269) = 4.98$, $p < .0001$). We see from our results (Figure 3.9) that for every increase in log(budget), the movie's runtime is expected to increase by about 3.56 minutes on average. However, an R^2 of .09 indicates that only 9% of the runtime variation is explained by the budget. Similar to our results from runtime and ratings, budget is a statistically significant factor, but not a strong predictor for a movie's runtime.



Dep. Variable:	runtime	R-squared:	0.086
Model:	OLS	Adj. R-squared:	0.083
Method:	Least Squares	F-statistic:	25.39
Date:	Sun, 07 Dec 2025	Prob (F-statistic):	8.60e-07
Time:	00:43:57	Log-Likelihood:	-1100.1
No. Observations:	271	AIC:	2204.
Df Residuals:	269	BIC:	2211.
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	50.6861	12.151	4.171	0.000	26.763	74.609
log_budget	3.5587	0.706	5.039	0.000	2.168	4.949

Figure 3.9 Log(Budget) v. Runtime Linear Model and Results

Conclusion

Summary

After completing all of our analysis, we were able to determine some of the answers to our research questions. We discovered that although sharing a high correlation, reviews tended to include more information that indicated a lower predicted rating score than the actual rating they assigned. In other words, the reviews people left behind on movies did not always reflect the actual rating they assigned to that movie, indicating that these reviews are valuable information to consider when judging how the audience perceived a movie. Humans tended to be more optimistic about their ratings than BERT predicted from their comments. We then discovered that higher budget movies tended to receive higher ratings and generate more revenue than lower budget movies. This makes sense, since movies with higher production costs are likely to have more money to spend on advertising as well, meaning that they can entice people into visiting the theaters through their advertising. As well as this, higher production costs will likely result in a more polished movie with better set design, scriptwriters, actors, and camera equipment. Therefore, audiences will likely respond better to the “high quality” movie than they would to

one without these features. Finally, we determined that movie longevity is a significant predictor of movie ratings and movie budgets, with longer movies resulting in higher ratings and higher budgets. We believe this makes sense since a longer screen time allows for a story to include more detail and finish, something that audiences likely appreciate. Furthermore, a longer runtime implies that there will be more scenes, sets, and dialogue, likely increasing the need to spend more money on actors, set, and prop rentals.

Limitations

Despite our significant findings, several areas of improvement still exist. For one, we limit our study to a subset of more recent movies but applying the same methodology to a larger set could prove useful in uncovering deeper patterns. Future research could also focus on other methods of collecting data, including exit polling from the theaters and other web sources with ratings and reviews. In regards to our analysis, other NLP methods could be used to assess the accuracy of reviews and ratings to see if a model can be closer in accurately predicting the two. Moreover, a key limitation with respect to the NLP is our assumption that a model is unbiased and accurately modeled human sentiment; pragmatically however, leveraging a state of the art model is the best quantitative proxy we have without extensive surveys. Additionally, when we took the log of the budgets to perform linear regression, we were forced to discard several movies with budgets of zero, limiting the scope of our data and potentially biasing the outcome of our regression.

References

Heckmann, C. (2022, February 13). *Film Budget and Production Expenses Explained*. StudioBinder. https://www.studiobinder.com/blog/production-budget/

Town, N. L. P. (2023). bert-base-multilingual-uncased-sentiment (Revision edd66ab). doi:10.57967/hf/1515

Figures and Tables

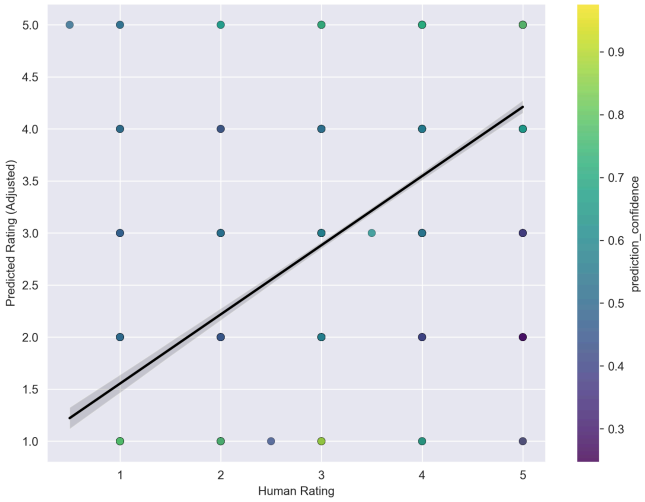


Figure 1.1: Regression of Human Rating vs Predicted Sentiment Rating

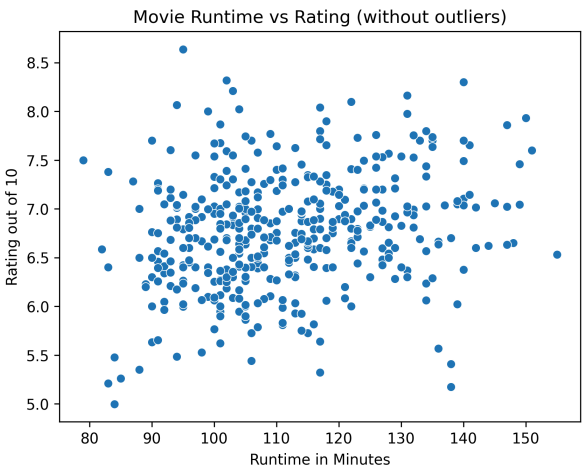


Figure 3.1: Scatterplot of Runtime v. Rating

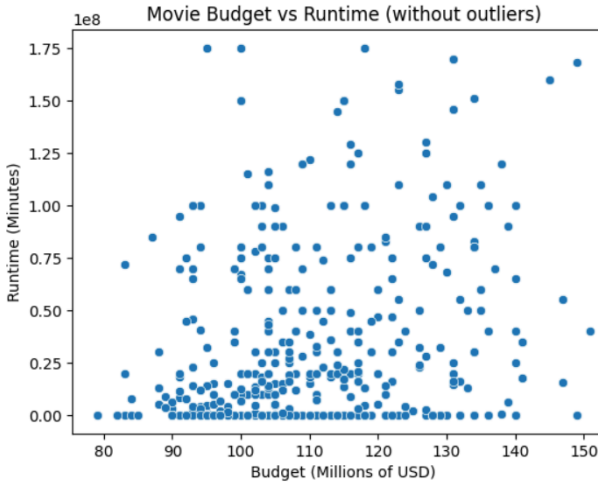


Figure 3.2: Scatterplot of Budget v. Runtime

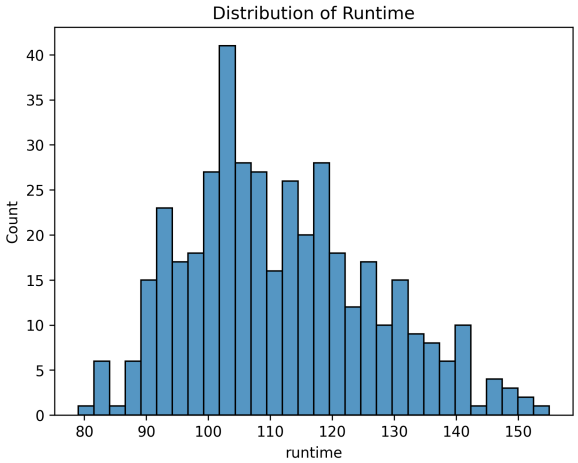


Figure 3.3: Distribution of Runtime

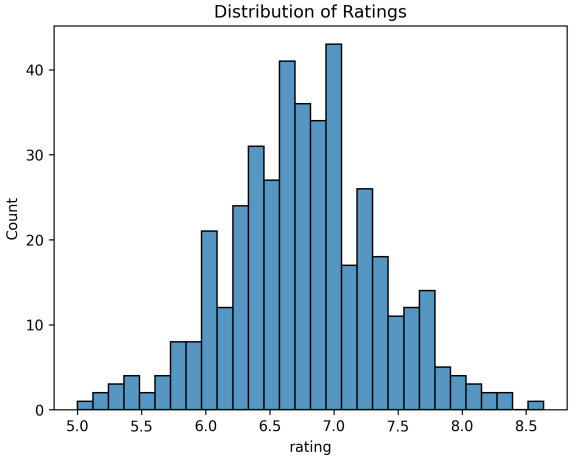


Figure 3.4: Distribution of Ratings

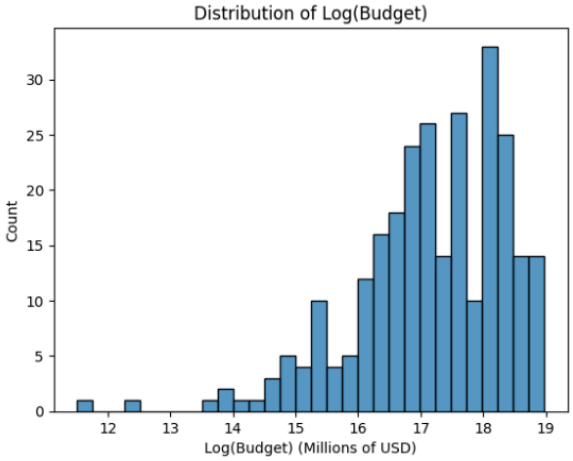


Figure 3.5: Distribution of Log(budget)

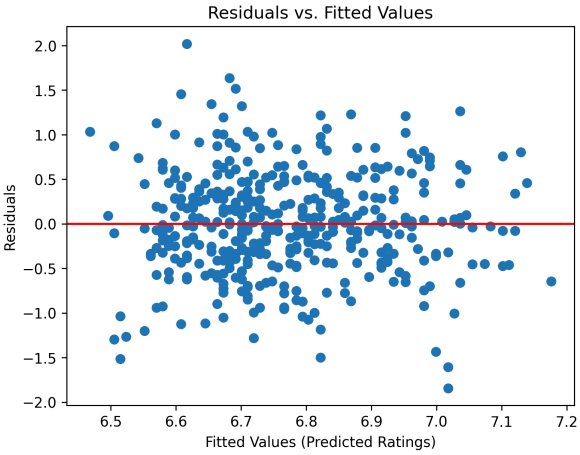


Figure 3.6: Residual Plot