Gordon Li
Professor Jacob Koehler
Data Bootcamp
16 December 2024
Word Count: 3498

Predicting Movie Lifetime Gross Revenue

Part 1: Introduction, Motivation, and Data Overview

Introduction

The modern movie industry is worth billions of dollars, with new movies coming out every week. Individuals enjoy movies through traditional theaters, streaming services such as Netflix, and occasionally through physical formats such as DVDs. The best movies can remain a part of a society for many years. Successful movie stakeholders capitalize on their success by issuing reruns, merchandise, sequels, and in rare instances, building entire cinematic universes. These can become incredibly profitable for everyone involved.

Motivation

Moviemakers want to ensure the maximum revenue possible. As such, I seek to build a model that can predict the Lifetime Gross Revenue of a movie. This model, if implemented, has many real world implications: directors and studios can use it to predict the lifetime gross profits of a movie after a release, enabling them to make early decisions on whether it is financially worthwhile to create a sequel. Investors can use these models to predict whether to invest in a movie or any projected sequels that get pitched to them.

Data Overview

The data contains information about the top 1000 movies by IMDB rating. The data was collected by Kaggle user Harshit Shankhdhar, who web scraped the information off of the IMDB website into a CSV file.

There are a total of 16 columns in the raw dataset:

1. Advertising poster link
2. Title
3. Release year
4. Certificate (rating of movie appropriateness) - G, PG, PG-13, etc.
5. Runtime (minutes)
6. Genre
7. IMDB rating
8. Movie overview/summary
9. Metascore (rating given by professional critics)
10. Director
11. Star Actor
12. Star Actor
13. Star Actor
14. Star Actor
15. Total Votes (on IMDB)
16. Gross Lifetime Revenue

The results of the data may be volatile due to certain metrics, such as ratings, being dependent on consumer tastes and sentiments which change over time. Additionally, the Gross column represents Lifetime Gross Revenue, which is a nominal value and does not take inflation into account.

The data will certainly require cleaning, as some rows lack key information. Additionally, the movie certificates use several different rating scales from different regulatory bodies.

Part 2: Data Cleaning

	Poster_Link	Series_Title	Released_Year	Certificate	Runtime	Genre	IMDB_Rating	Overview	Meta_score	Director	Star1	Star2	Star3	Star4	No_of_Votes	Gross
0	https://m.media-amazon.com/images/M/MV5BMDFkYT...	The Shawshank Redemption	1994	A	142 min	Drama	9.3	Two imprisoned men bond over a number of years...	80.0	Frank Darabont	Tim Robbins	Morgan Freeman	Bob Gunton	William Sadler	2343110	28,341,469
1	https://m.media-amazon.com/images/M/MV5BM2MyNj...	The Godfather	1972	A	175 min	Crime, Drama	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	Marlon Brando	Al Pacino	James Caan	Diane Keaton	1620367	134,966,411
2	https://m.media-amazon.com/images/M/MV5BMTMxNT...	The Dark Knight	2008	UA	152 min	Action, Crime, Drama	9.0	When the menace known as the Joker wreaks havo...	84.0	Christopher Nolan	Christian Bale	Heath Ledger	Aaron Eckhart	Michael Caine	2303232	534,858,444
3	https://m.media-amazon.com/images/M/MV5BMWMwMG...	The Godfather: Part II	1974	A	202 min	Crime, Drama	9.0	The early life and career of Vito Corleone in ...	90.0	Francis Ford Coppola	Al Pacino	Robert De Niro	Robert Duvall	Diane Keaton	1129952	57,300,000
4	https://m.media-amazon.com/images/M/MV5BMWU4N2...	12 Angry Men	1957	U	96 min	Crime, Drama	9.0	A jury holdout attempts to prevent a miscarria...	96.0	Sidney Lumet	Henry Fonda	Lee J. Cobb	Martin Balsam	John Fiedler	689845	4,360,000

First few rows of data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 16 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Poster_Link    1000 non-null   object 
 1   Series_Title   1000 non-null   object 
 2   Released_Year  1000 non-null   object 
 3   Certificate    899 non-null    object 
 4   Runtime        1000 non-null   object 
 5   Genre          1000 non-null   object 
 6   IMDB_Rating    1000 non-null   float64
 7   Overview       1000 non-null   object 
 8   Meta_score     843 non-null    float64
 9   Director       1000 non-null   object 
 10  Star1          1000 non-null   object 
 11  Star2          1000 non-null   object 
 12  Star3          1000 non-null   object 
 13  Star4          1000 non-null   object 
 14  No_of_Votes    1000 non-null   int64  
 15  Gross          831 non-null    object 
dtypes: float64(2), int64(1), object(13)
memory usage: 125.1+ KB
```

Summary of data

We see above that there is quite a lot of missing data in the Certificate, Meta_score, and Gross columns. Additionally, Released_Year should be converted to an integer, and Runtime and Gross should be integers. We correct these mistakes. We also combine the actor columns and split the genre columns and drop rows with null values.

The Certificate ratings column uses multiple rating scales. Using the IMBD reference and this government article and blog on Indian rating scales, I will convert ratings into the American system (G, PG, PG-13, R) as follows:

U = G
GP and TV-PG = PG
UA and U/A = PG-13
A = R

I will remove movies rated "Passed" and "Approved", as they fall under the old requirements of Hayes's Law.

Part 3: Data Investigation and Exploratory Data Analysis

Preliminary Investigations - Numerical Variables

	Poster_Link	Series_Title	Released_Year	Certificate	Runtime	Genre	Genre 1	Genre 2	Genre 3	IMDB_Rating	Overview	Meta_score	Director	Stars	Star1	Star2	Star3	Star4	No_of_Votes	Gross
0	https://m.media-amazon.com/images/M/MV5BMDFkYT...	The Shawshank Redemption	1994	R	142	Drama	Drama	None	None	9.3	Two imprisoned men bond over a number of years...	80.0	Frank Darabont	Tim Robbins, Morgan Freeman, Bob Gunton, Willi...	Tim Robbins	Morgan Freeman	Bob Gunton	William Sadler	2343110	28341469
1	https://m.media-amazon.com/images/M/MV5BM2MyNj...	The Godfather	1972	R	175	Crime, Drama	Crime	Drama	None	9.2	An organized crime dynasty's aging patriarch t...	100.0	Francis Ford Coppola	Marlon Brando, Al Pacino, James Caan, Diane Ke...	Marlon Brando	Al Pacino	James Caan	Diane Keaton	1620367	134966411

First few rows of cleaned data

	Released_Year	Runtime	IMDB_Rating	Meta_score	No_of_Votes	Gross
count	699.000000	699.000000	699.000000	699.000000	6.990000e+02	6.990000e+02
mean	1996.776824	123.695279	7.935050	76.937053	3.618417e+05	8.002793e+07
std	17.305745	26.036730	0.294922	12.364986	3.554662e+05	1.157291e+08
min	1930.000000	72.000000	7.600000	28.000000	2.522900e+04	1.305000e+03
25%	1987.500000	104.000000	7.700000	69.000000	9.850300e+04	6.727491e+06
50%	2001.000000	120.000000	7.900000	78.000000	2.415750e+05	3.590000e+07
75%	2010.000000	136.500000	8.100000	86.000000	5.159115e+05	1.066073e+08
max	2019.000000	238.000000	9.300000	100.000000	2.343110e+06	9.366622e+08

Numerical Data - Summary Statistics

The cleaned dataset now has 699 movies. The movie release years range from 1930 to 2019. Runtimes range from just 72 minutes to 238 minutes, with the average being around 2 hours. IMDB ratings range from 7.6 to 9.3, with the average being a 7.9. The Metacritic scores range from from 28 to a perfect 100, with the average around a 77. The number of votes for a movie range from just over 25,000 to 2.3 million votes, with the average number of votes at 362,000.

Preliminary Investigations - Categorical Variables

Certificate	count
R	305
G	192
PG-13	181
PG	21

Certificate Count

	0
Drama	489.0
Adventure	160.0
Comedy	159.0
Crime	141.0
Action	138.0
Thriller	97.0
Biography	88.0
Romance	84.0
Mystery	68.0
Animation	63.0
Fantasy	55.0
Sci-Fi	55.0
Family	43.0
History	37.0
Music	27.0
War	27.0
Sport	17.0
Horror	16.0
Western	13.0
Musical	11.0
Film-Noir	3.0

Genre Count

Director	count
Steven Spielberg	13
Martin Scorsese	10
Clint Eastwood	8
Christopher Nolan	8
Quentin Tarantino	8
...	...
Alan J. Pakula	1
Alejandro Jodorowsky	1
Luis Buñuel	1
Sam Peckinpah	1
Richard Lester	1

Director Count

	0
Robert De Niro	16.0
Tom Hanks	14.0
Al Pacino	13.0
Brad Pitt	12.0
Clint Eastwood	11.0
...	...
Jacki Weaver	1.0
Jack Thibeau	1.0
Jack Reynor	1.0
Jack McBrayer	1.0
Éric Toledano	1.0

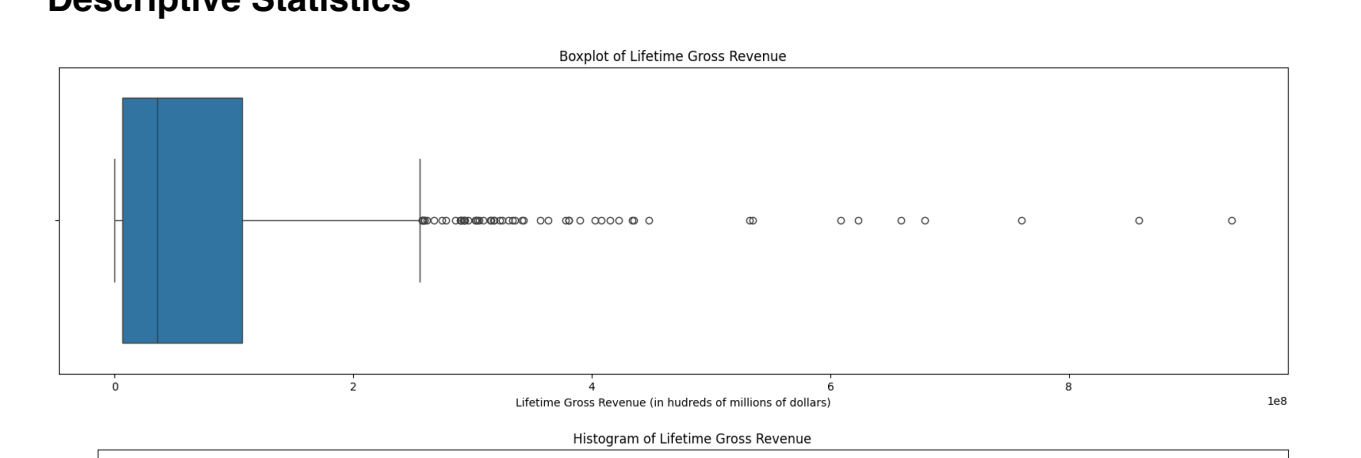
1870 rows × 1 columns

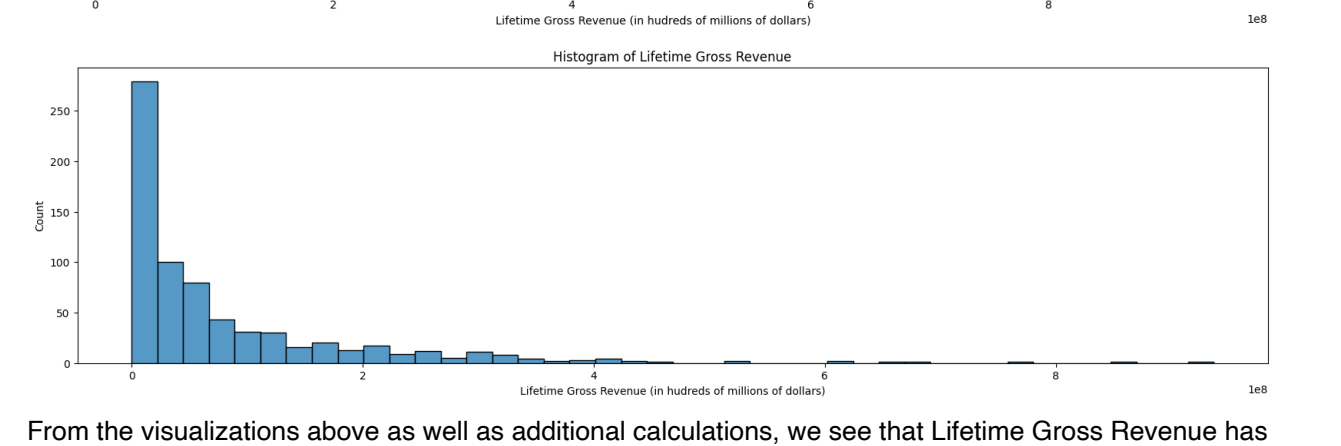
Stars Count

The categorical variables include the movie Certificate, Genre(s), Director, and movie stars. The four certificates, G, PG, PG-13, and R, have counts of 192, 21, 181, and 305, respectively. The most common certificate is R, and the least common is PG. There are 21 different genres, with by far the most common being Drama, with 489 movies in this category. The movies represent 395 directors, with Stephen Spielberg directing the most movies, at 13. There are 1870 movie stars named in the data, with Robert De Niro being named the most at 16 times, then Tom Hanks at 14, and Al Pacino at 13.

Exploratory Data Analysis

Descriptive Statistics

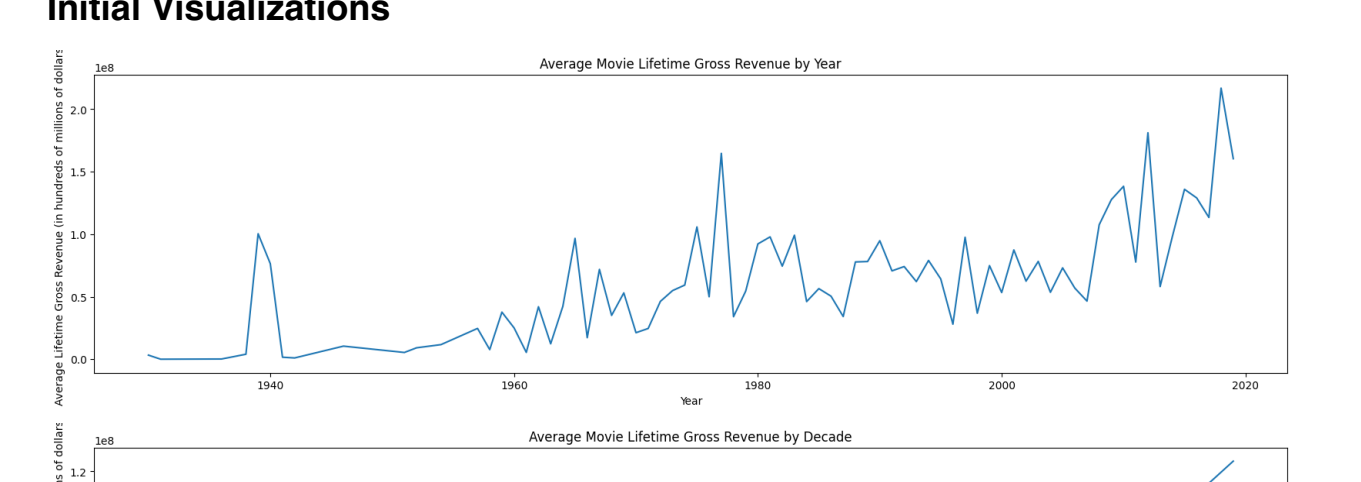


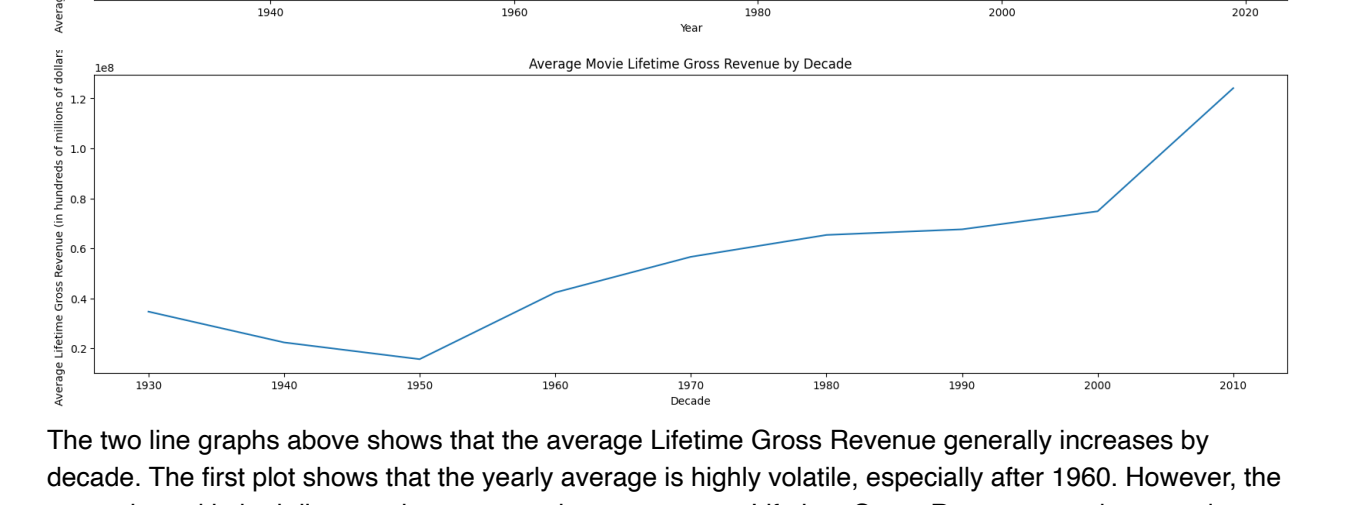


From the visualizations above as well as additional calculations, we see that Lifetime Gross Revenue has an average of around 80 million dollars, a median of around 35 million dollars, a standard deviation of over 115 million dollars, and ranges from just 1305 dollars to 936,662,225 dollars.

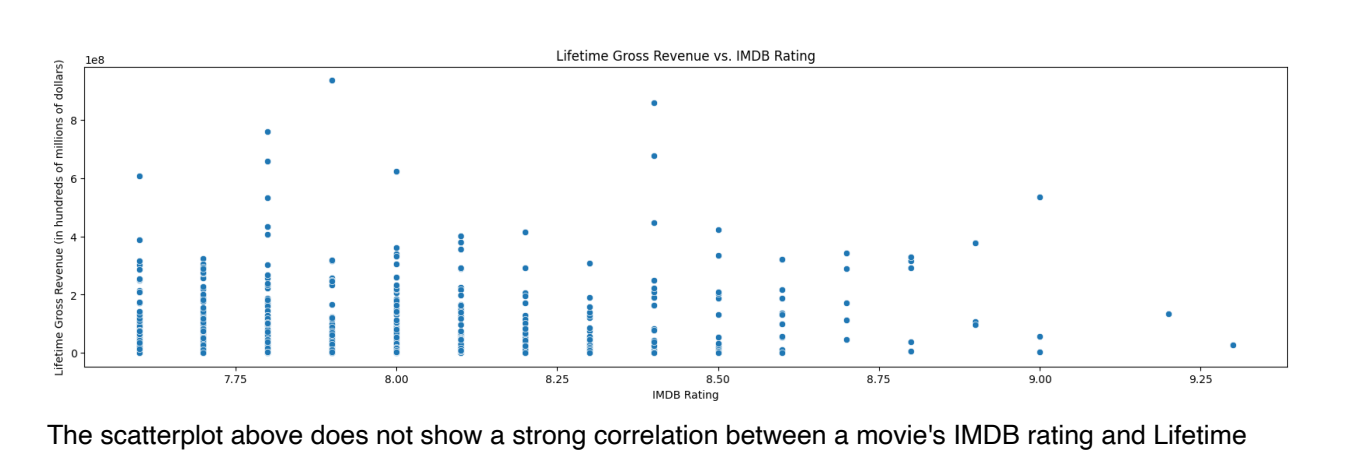
The data is heavily skewed, with many high outliers. This could partly result from the fact that relatively few movies make large amounts of money while most movies make very little money. This could also be the result of using Nominal Revenue data instead of using Real Revenue (i.e. adjusted for inflation).

Initial Visualizations

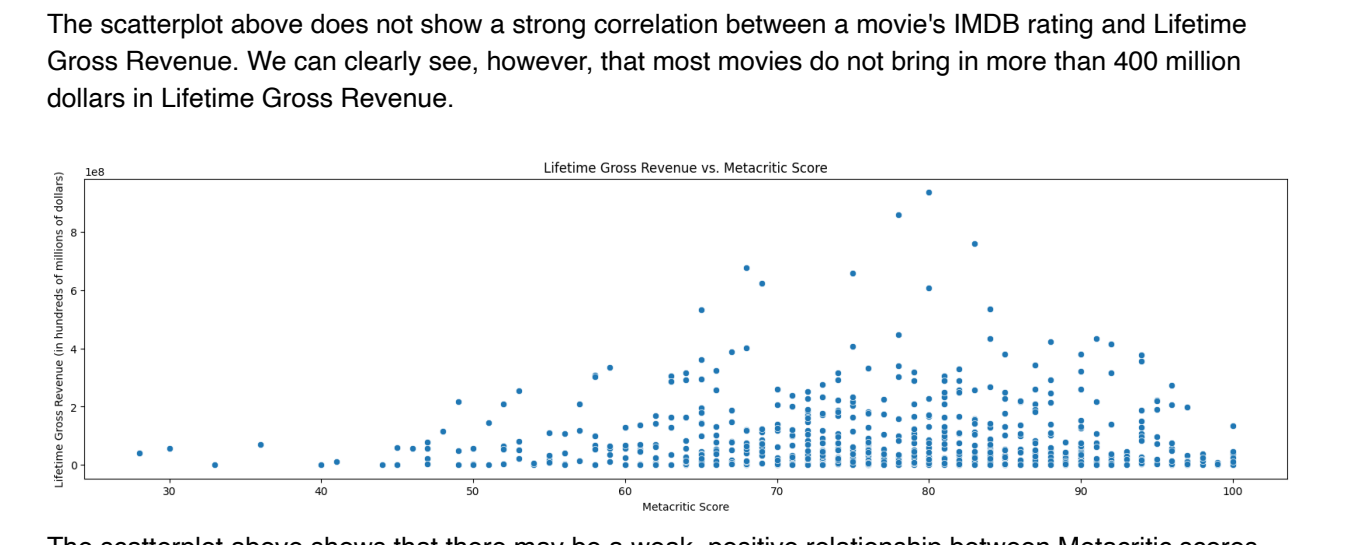




The two line graphs above shows that the average Lifetime Gross Revenue generally increases by decade. The first plot shows that the yearly average is highly volatile, especially after 1960. However, the general trend in both line graphs suggests that on average, Lifetime Gross Revenue trends upward over time.

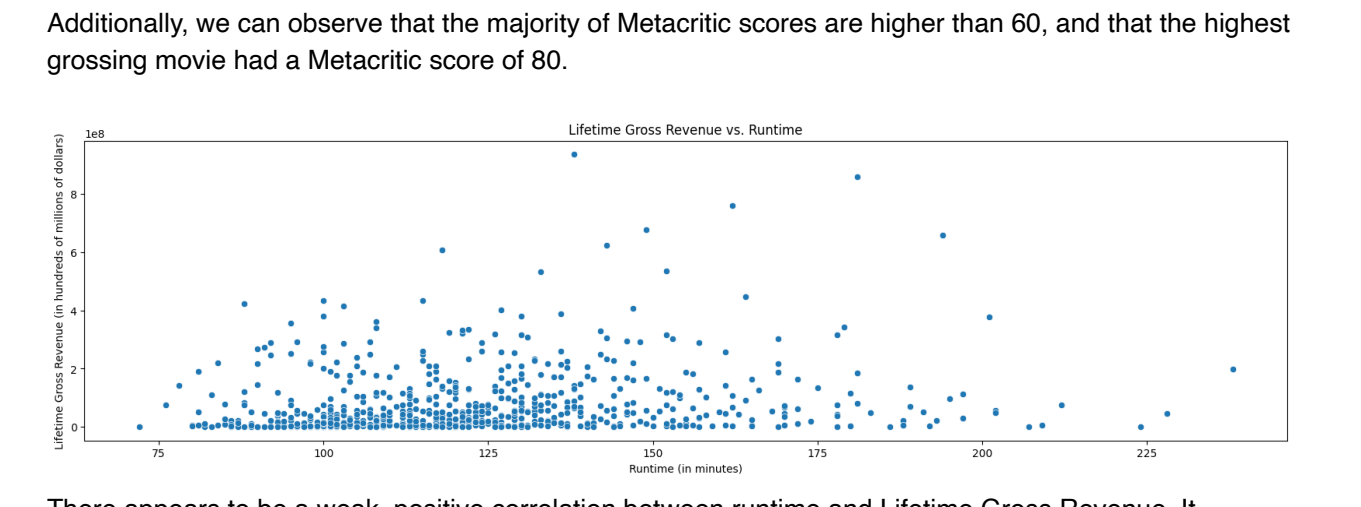


The scatterplot above does not show a strong correlation between a movie's IMDB rating and Lifetime Gross Revenue. We can clearly see, however, that most movies do not bring in more than 400 million dollars in Lifetime Gross Revenue.



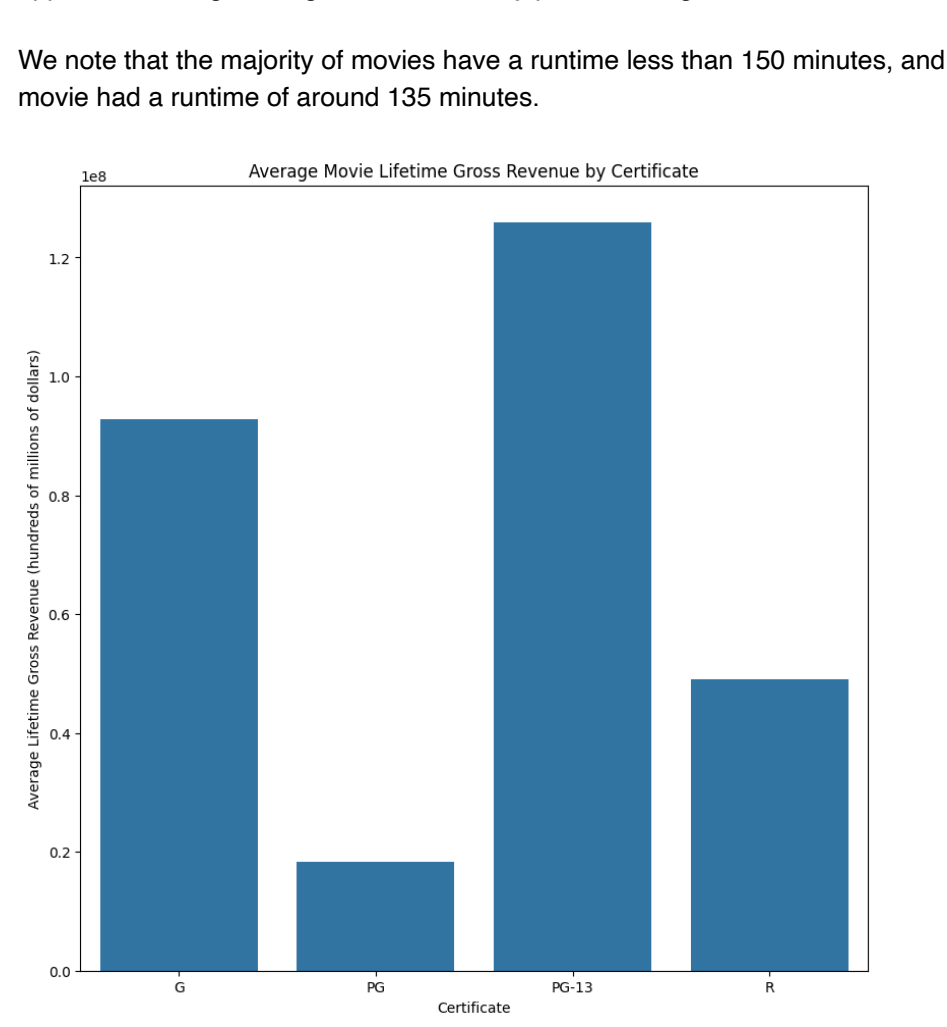
The scatterplot above shows that there may be a weak, positive relationship between Metacritic scores and a movie's Lifetime Gross Revenue. Movies with higher critic scores appear to have a higher Lifetime Gross Revenue.

Additionally, we can observe that the majority of Metacritic scores are higher than 60, and that the highest grossing movie had a Metacritic score of 80.

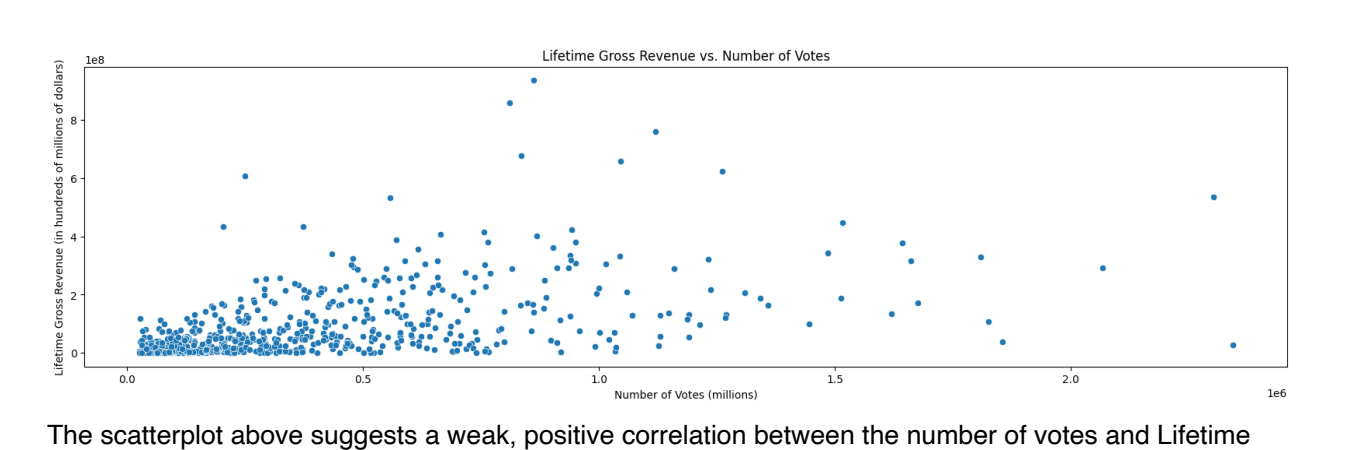


There appears to be a weak, positive correlation between runtime and Lifetime Gross Revenue. It appears as though a longer runtime weakly predicts a higher Lifetime Gross Revenue.

We note that the majority of movies have a runtime less than 150 minutes, and that the highest grossing movie had a runtime of around 135 minutes.



We see that PG-13 movies, on average, have the highest average Lifetime Gross Revenue. PG movies have the lowest. Interestingly, there does not seem to be a correlation between rating level and average Lifetime Gross Revenue.

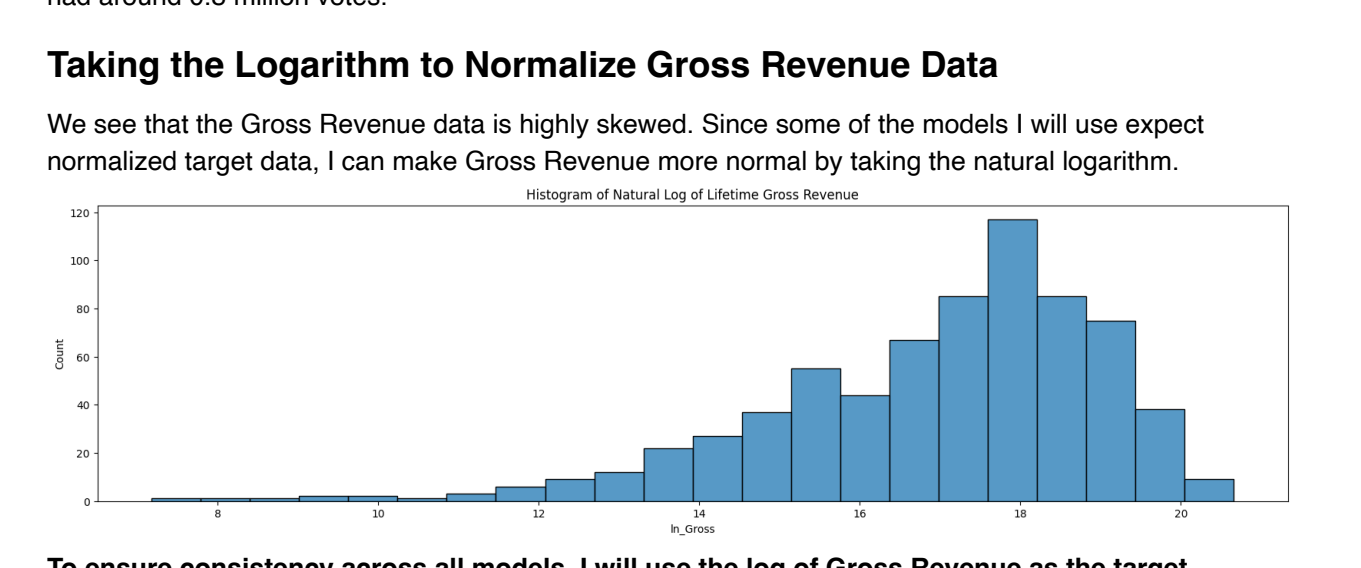


The scatterplot above suggests a weak, positive correlation between the number of votes and Lifetime Gross Revenue. As the number of votes increases, one can weakly infer a higher Lifetime Gross Revenue. This could be due to the fact that a higher number of votes means greater interest in the movie, which means more ticket sales and more revenue.

We note that the majority of movies have fewer than 0.75 million votes, and the highest grossing movie had around 0.8 million votes.

Taking the Logarithm to Normalize Gross Revenue Data

We see that the Gross Revenue data is highly skewed. Since some of the models I will use expect normalized target data, I can make Gross Revenue more normal by taking the natural logarithm.



To ensure consistency across all models, I will use the log of Gross Revenue as the target variable from now on.

A Note on Predictor Variables

In some of the models below, I will use the IMDB Rating, Metascore, and total number of votes as predictor variables. While this is not ideal as these values manifest after a movie's release, I justify their use in the following ways:

1. Timing of Metascore: The Metascore represents the ratings of professional movie raters, and they tend to form their opinions on the opening night. As such, the Metascore value will remain fairly fixed soon after the movie's release, well before a movie reaches its Lifetime Gross Revenue.
2. Timing of IMDB Ratings and Total Votes: Previous research shows that IMDB ratings tend not to fluctuate much after the first year, and the total number of votes for a given movie tends to follow a logistic pattern that approaches the upper asymptote after about 50 days. Therefore, as long as the model is used one year after the movie release (see point 4), the data is a good approximation for what would be available to other users of my models.
3. Data Limitations: IMDB does not easily provide information on how the ratings and votes for a movie change over time.
4. Intent of Models: In the introduction, I state that the goal of these models is to help investors, directors, and studios decide whether they should invest/commit to sequels, further releases, and franchising opportunities. These decisions are made after observing the initial performance of a movie (after a few months to a year post release).

As such, the models generated with these columns as predictors will still be useful for predicting Lifetime Gross Revenue since my data approximates the values that would be available to people using the models for their intended purpose.

Part 4: Models and Interpretation

To predict Lifetime Gross Revenue, I will generate four regression models to see which one performs the best. I will also create a fifth model involving a Neural Network to predict whether a movie will reach a certain Lifetime Gross Revenue given its overview (synopsis).

For each model, I will divide my existing data into training and testing datasets in an 80-20 train-test split, in line with industry-standard data analysis practices. This means that 80% of the data will be used to train the model, and 20% will be used to test the model.

Baseline Model

I will evaluate the success of my models by comparing against a baseline. I choose to compare the Mean Absolute Percentage Error (MAPE) because it's a fairly easy to understand metric that can be used across various models.

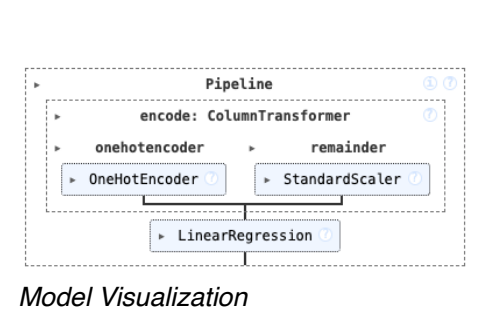
My baseline will be the mean of the natural log of the Lifetime Gross Revenues in my dataset, which I calculated previously.

My calculations show that the baseline MAPE is 0.104.

Model 1: Multiple Linear Regression Model

My first model is a Multiple Linear Regression model. I choose to use this model because it is the most straightforward and easiest to understand. Using this model, we can identify the effects of individual predictors on Gross Revenue, as well as their combined effects.

Model Output



Model Visualization

	0
No_of_Votes	1.684543
Runtime	0.278400
Meta_score	0.217226
Released_Year	-0.071009
Certificate_PG-13	-0.529368
Certificate_PG	-0.965169
Certificate_R	-1.007764
IMDB_Rating	-1.106580

Coefficients

	0
No_of_Votes	2.077449
IMDB_Rating	0.859913
Certificate	0.085211
Runtime	0.065540
Meta_score	0.026095
Released_Year	0.003529

Feature Importance

The y-intercept is 17.541467790523246
The MAPE for the training data is 0.07494305063647635
The MAPE for the testing data is 0.06972479908372053

Analysis

My Multiple Linear Regression model performed better than the baseline. Both the test and training data performed better than the baseline, since the test MAPE of 0.0697 and training MAPE of 0.0749 are both smaller than the baseline MAPE of 0.104. This is an indication that this model accounted for the impacts of Metacritic score, runtime, votes, released year, IMDB rating, and certificate on the log of Lifetime Gross Revenue, which the baseline obviously could not.

Metacritic score, runtime, and number of votes all had a positive impact on the log of Lifetime Gross Revenue. Strangely, the released year and number of votes seems to negatively correlate with the log of Lifetime Gross Revenue. This seems counterintuitive because I would expect more votes to equal more attention and popularity, and also for more recent movies to benefit from a greater demand from entertainment in recent years.

The most important input was Number of Votes, followed by IMDB rating. The released year appears to be the least important input in this model.

Model 2: Quadratic Polynomial Model

We can build on the Linear Model above to incorporate quadratic coefficients. I choose to do this because a quadratic model may better capture any possible curved trends or more complex patterns in the data.

Model Output

```
LinearRegression
LinearRegression()
```

Model Visualization

	0
Certificate_PG^2	2.729244e+23
Certificate_PG	2.729244e+23
No_of_Votes	2.677658e+00
No_of_Votes^2	6.720207e-01
IMDB_Rating	4.743934e-01
IMDB_Rating No_of_Votes	1.535650e-01
Released_Year Runtime	9.358355e-02
Runtime^2	4.736568e-02
Certificate_R Runtime	4.730596e-02
Released_Year Meta_score	2.895729e-02
Certificate_PG-13 Runtime	2.798123e-02
Certificate_R^2	2.660088e-02
Certificate_R	2.582696e-02
Released_Year No_of_Votes	2.290270e-02
Runtime	1.628570e-02
Certificate_PG-13 No_of_Votes	1.494167e-02
IMDB_Rating Meta_score	1.410635e-02
Certificate_R Meta_score	1.061924e-02
Certificate_R Released_Year	8.293058e-03
Released_Year^2	7.540975e-03
Released_Year	7.460465e-03
Runtime IMDB_Rating	6.710901e-03
Meta_score^2	6.300030e-03
Runtime No_of_Votes	6.131523e-03
Certificate_PG IMDB_Rating	5.890715e-03
Released_Year IMDB_Rating	5.116767e-03
Certificate_PG-13 Meta_score	4.566950e-03
Certificate_PG Released_Year	3.715989e-03
Certificate_PG-13 IMDB_Rating	2.791382e-03
Meta_score No_of_Votes	2.736478e-03
Meta_score	1.581376e-03
Certificate_R IMDB_Rating	7.916780e-04
Certificate_PG-13 Released_Year	7.123942e-04
Runtime Meta_score	6.026952e-04
IMDB_Rating^2	6.097305e-05
Certificate_PG-13 Certificate_R	0.000000e+00
Certificate_PG Certificate_R	0.000000e+00
Certificate_PG Certificate_PG-13	0.000000e+00
Certificate_PG Meta_score	-8.609630e-05
Certificate_PG Runtime	-2.336327e-04
Certificate_R No_of_Votes	-1.225240e-03
Certificate_PG No_of_Votes	-3.802928e-03
Certificate_PG-13	-3.980273e-03
Certificate_PG-13^2	-3.994029e-03

Feature Importance

There are 44 coefficients in the model
The MAPE for the training data is 0.06536089199942088
The MAPE for the testing data is 0.05844636246110065

Analysis

The Quadratic model performed better than the baseline and the linear regression model for both the training and testing data. This is shown by the fact that the Quadratic model had a training MAPE of 0.065 and a testing MAPE of 0.058. This is better than the Linear model, which had a training MAPE of 0.0749 and a test MAPE of 0.0697. Of course, this is all better than the baseline MAPE of 0.104.

We see that the model has 44 features. The most important features are the Certificate_PG, followed by Number of Votes. Number of Votes remains influential, but clearly is not as impactful as in the linear model.

Model 3: Random Forest Model

Seeing as the non-linear Quadratic model performed better than the Linear one, I wanted to explore models that might further investigate non-linear relationships in the data. Since an ensemble model like Random Forests combines multiple decision trees, I hope to see if I can further improve my predictive abilities by finding the "best" model.

Model Output

	0
No_of_Votes	1.353544
Released_Year	0.210148
IMDB_Rating	0.147631
Certificate	0.074665
Runtime	0.056902
Meta_score	0.046463

Feature Importance

Best parameters: {'model__max_depth': 8, 'model__n_estimators': 100}
The MAPE for the training data is 0.035942091986670394
The MAPE for the testing data is 0.060018968682073584

Analysis

The Random Forest model determined that the best decision tree has a maximum depth of 8 and 100 estimators. The selected model performed better than the baseline, since it had a training MAPE of 0.035 and a testing MAPE of 0.06, significantly better than the baseline MAPE of 0.104. While the model performed better in training compared to both the linear or quadratic models, the model did slightly worse than the quadratic model for the test data.

The most important features in this model were the Number of Votes, with Released Year coming in second. This is similar to the Linear model, where Number of Votes was the most important feature.

Model 4: K-Nearest Neighbors Model

To further investigate the normalized data and other possible non-linear patterns, I choose to build a KNN model. A KNN model makes predictions based on the surrounding figures. Therefore, if there are any clustering or local patterns that heavily influence Lifetime Gross Revenue, this model should identify them and make good predictions. By using a grid search, we can further identify how many neighbors (i.e. which "K" value) is optimal.

Model Output

	0
No_of_Votes	0.931499
Released_Year	0.204127
IMDB_Rating	0.147830
Runtime	0.129212
Certificate	0.103346
Meta_score	0.058732

Feature Importance

Best parameters: {'model__n_neighbors': 5}
The MAPE for the training data is 0.06294402985266428
The MAPE for the testing data is 0.06422136677865391

Analysis

Out of the different K-values tested, the best model considers 5 of the nearest neighbors. This model performed better than the baseline for both training and testing data, since it has a training MAPE of 0.063 and a testing MAPE of 0.064. This is better than the baseline MAPE of 0.104.

The model performed better than the quadratic and linear models on the training data, as its MAPE of 0.063 is smaller than the Quadratic and Linear training MAPEs. However, while the model performed better than the linear model on the testing data, it failed to outperform the quadratic model. This suggests that there is not significant local clustering in the data. The model also failed to outperform the Random Forest model on both training and testing data.

The most significant feature was the Number of Votes, which is consistent with the Linear and Random Forest models.

Model 5: Neural Network for Binary Prediction

For my last model, I thought it would be interesting to see if it would be possible to use a neural network to do text classification. Some investors, upon receiving a synopsis during a movie pitch, may want to determine whether the movie is likely to reach a certain Lifetime Gross Revenue. This model seeks to determine whether a model is able to predict whether a movie will reach the average Lifetime Gross Revenue based on its synopsis.

We create a new dataframe with a binary classification column, depending on whether the Lifetime Gross Revenue is higher or lower than the average. We will need to create a new baseline and evaluation metric.

Baseline and Evaluation Metric

We use Accuracy as our evaluation metric, which is defined as (True Positives + True Negatives) / Total. We choose this because both true positives and true negatives matter: investors cannot afford to invest in unprofitable movies, but they also cannot afford to miss out on good ideas.

We calculate the baseline accuracy with the assumption that all movies fail to achieve the mean Lifetime Gross Revenue. The baseline accuracy is 0.4857142857142857

Model Accuracy - Test Data

We calculate the baseline accuracy using the Test Data and the Prediction output. The model's accuracy is 0.5.

As we observe, the model is a bit more accurate than the baseline, indicating that it may be moderately useful in assessing whether a movie's log of Gross Revenue will be higher than the average.

Part 5: Summary of Findings

My analysis consisted of five models: a Linear model, Quadratic model, Random Forest model, KNN model, and a Neural Network for binary predictions. All models performed better than the baseline, highlighting their usefulness in potentially helping investors and analysts project the log Lifetime Gross Revenue. In addition to directly predicting the lifetime revenue of the movie in question, this information may also be used to investigate the feasibility and profitability of reruns, sequels, and other related revenue streams.

The following are a few key findings and takeaways:

1. The four regression models listed from best to worst performance are as follows: Quadratic, Random Forest, KNN, and Linear.
2. The Quadratic model is quite robust, performing better than the other models and significantly better than the baseline with a testing MAPE of 0.058. This could be due to the Quadratic model's ability to account for non-linear trends and account for interaction effects by adding many additional terms. This highlights its effectiveness and usefulness for making predictions about Lifetime Gross Revenue.
3. The number of votes was the most impactful feature of the Linear, Random Forest, and KNN models, and a close second place for the Quadratic model. This could be due to a higher number of votes implying a higher movie popularity, which implies higher interest and viewership and thus higher revenues.
4. The Neural Network performed better than the baseline, implying that it is useful for predicting if a movie will make more than the average log Gross Revenue. This is likely due to successful movies in the dataset having similar themes (adventure, drama, etc), meaning the model could identify keywords that could distinguish between the above and below average groups.

Overall, the Quadratic model is likely the best model to use for predicting log Lifetime Gross Revenue. The number of votes remains an influential factor regardless of model, likely due to it acting as an indicator for overall movie interest and popularity. These takeaways offer a detailed view into the many factors that affect a movie's potential ability to make money, highlighting the complex and highly difficult task that studios, investors, directors, and other stakeholders face when deciding on the next big film.

Part 6: Discussion and Possible Next Steps

To strengthen the predictions and usefulness of these models, I believe that future researchers and analysts could implement the following:

1. **Adjust data for inflation:** I think one major drawback and a major cause of skewness in my data was the lack of real values for Gross Revenue. With inflation-adjusted data, I believe my model would likely be more normal and thus require less adjusting, resulting in easier comparisons and a better model overall.
2. **Add in information about production costs:** my data did not include the costs of production. I believe that if I had this data, I could have used Revenue per Dollar invested as my target instead of Gross Revenue. This metric is more appropriate for investors and studios who are looking to decide how much to invest, and could also avoid some of the concerns regarding inflation since taking the ratio would allow us to cancel out the inflation factor.
3. **Add in time-series information:** I would be really interested in seeing how the model projections change as the input values, such as total votes and rating, change over time. Again, this would be very valuable to investors who have to make a decision quickly, but want to see a range of projected revenues given the currently available data and possible future trends.
4. **Add in economic data:** Movies are luxury goods, and one of the things that people cut out when money gets tight are subscription services and movie dates. As such, when the movie is released relative to the business/economic cycle can greatly affect its initial popularity, which has lasting consequences for its future legacy.

Integrating these factors would allow me to further hone my analysis and provide even more robust predictions and results, making my models even more valuable to investors, producers, analysts, and movie aficionados.