Abstract:

Using a dataset found on Kaggle[1] we trained a multiple linear regression model to predict the success of a movie as measured by generated revenue based on factors such as budget, director gender, and release year. The original data contained two datasets: one with information about movies (budget, popularity, revenue, title, vote statistive, the ID of the movie director, and the year, month and day of the week of the movie's release) the other about movie directors (director name, director ID, and the director's gender were included). We did not use the popularity variable in our regression since the dataset made it unclear what "popularity" was exactly or how this metric was calculated.

After cleaning the data by removing any movies that lacked information about their budget and revenue we ran two regression models. The first gave us an R^2 value of 0.63 and an MSE of 1.57e+16. Although this performance was pretty solid (above 0.5 R^2 is usually ideal) we were worried that by including every variable from the data as part of our regression we were overfitting the model. The regression was rerun with fewer variables and we ended up with an R^2 of 0.66 and MSE of 1.43e+16.

The regression coefficients showed us that both the movie's budget and vote count were highly correlated to movie success (revenue). Neither of these findings was especially surprising. A higher budget means more resources which, generally speaking, leads to a higher quality product and therefore higher revenue, and vice versa although this is not always the case. Having more votes means a movie had more audience engagement, more people watched the movie, thus

[1] Ck, N. (2024, September 30). *Movies and directors: Dataset for film analytics*. Kaggle. https://www.kaggle.com/datasets/nayanack/movies-and-directors-dataset-for-film-analytics

also increasing revenue. Release timing, especially around holidays, also had a positive effect on a movie’s success.

Predictors of Movie Success

Understanding what drives a movie's financial success is a critical challenge for the film industry. By identifying key predictors of revenue, studios can make informed decisions about budgeting, marketing, and release strategies. These predictors like release date are even more relevant now due to the rise of streaming services, as studios rely on high success of films and large revenues to compete with the online only entertainment model of streaming. By finding what factors are most related to revenue, studios can change their strategy and allocate resources to maximize revenue. This analysis focuses on predicting movie revenue using a streamlined set of predictors: budget, audience ratings (vote count and average), director gender, and release date.

Our goal is to develop a regression model that evaluates how these factors contribute to box office performance. Budget reflects the scale of production and marketing efforts, while vote count and average represent audience reception and critical acclaim. Director gender provides insight into demographic influences, and release date captures seasonal and temporal trends that may affect audience turnout.

Through this approach, we aim to highlight the relative importance of these predictors in driving revenue, offering actionable insights for optimizing future film projects. By focusing on these targeted variables, we provide a practical framework for understanding the interplay between production decisions and financial outcomes in the movie industry.

We found our dataset on Kaggle and the data had already been cleaned. The data contains two csv files, one with information about various movies and the other with information about movie directors. The movie dataset has columns for budget, popularity, revenue, title, vote statistive, the ID of the movie director (foreign key to the director dataset) and the year, month and day of the week of the movie's release. The director dataset has three columns: the director name, director ID, and the director's gender.

Our goal is to analyze what factors correlate with a movie's success, with our metric for success being revenue. We will use this data to see what movies were successful, when they were released, and who directed the movies. We would like to look at some factors outside of just what is in our dataset, however, finding additional data that has what additional metrics we would be interested in such as leading actors or composers and has that information for a large enough portion of the movies in the data set may prove difficult.

Methods:

We decided that for the data we had and we were trying to predict using a regression model would be the "best fit". More specifically we decided on a multiple, or multivariate, linear regression since the goal of our analysis is to predict movie success based on a variety of known factors (such as movie budget, director, etc). Our metric for the success of our model will be whether it correctly predicts how much revenue a film will make. A successful film in our analysis will be based on the financial success, so our model's approach will be to predict the revenue of a film based on its characteristics.

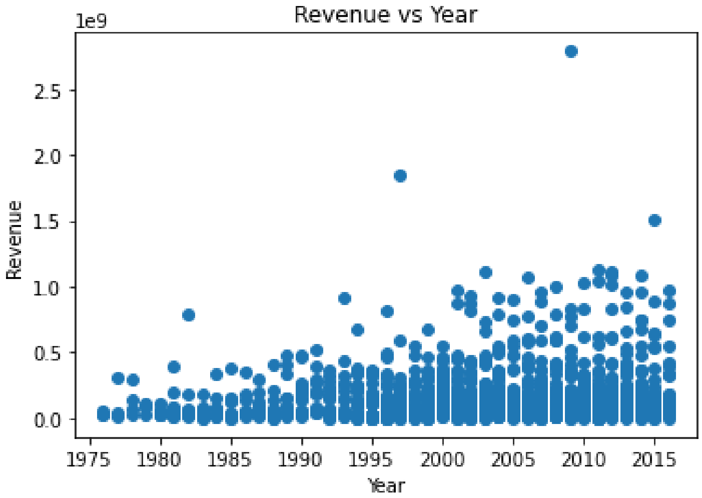
The main issue we anticipate is with data quality. Missing and incomplete data in important columns like budget and revenue could lead to major issues. One possible way to deal with them could be to make reasonable estimates for missing columns by using averages or predictive models. Fortunately, because the data is already in a usable format we don't have to do much in regards to feature engineering just taking the relevant information from the directors dataset and adding to the data from the movies dataset and removing any observations, movies, that are missing crucial information.

Finally, our results will be communicated visually via graphs and charts to show the data and the model's predictions as well as by presenting the R^2 and regression coefficient values.

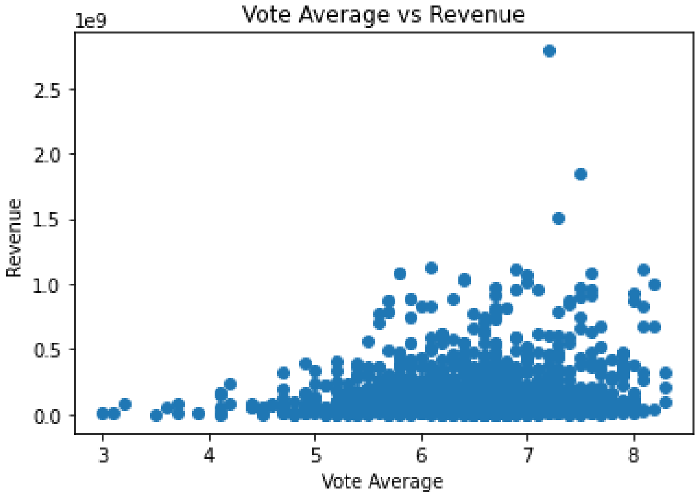
We started by removing movies in the dataset that had no revenue and/or no budget as it is highly unlikely that a movie cost $0 to make and, being erroneous observations, would skew our model. Since we are trying to predict a movie's success based on its revenue movie's without their listed revenue are unhelpful for the training and testing of the model. There was an index/primary key column in the dataset that we removed since it is not actually "data", and since the data itself was split into two files, we took the information we are interested in, director gender, from the directors.csv and added it to the data taken from movies.

Our initial regression used almost every column of data that we were given in the original dataset as a variable for the regression. The resulting R^2 value was 0.63, which is a solid performance considering that generally above 0.5 is ideal. The MSE was very large (1.57e+16), this was perhaps in part because of outliers or just large differences between big blockbuster-type movies versus relatively smaller scale movies. The budget seems to be highly correlated to the revenue which is something that we expected. Male directors on average make movies that perform better. The months of November and December do well in terms of revenue, likely because it is around the time of many holidays, and for students breaks, so there is more free time for people to spend at the movies. Certain days of the week did better than others as well. For example, Tuesday did better than Saturday, which at first seems arbitrary but is perhaps because of certain theaters offering deals and sales on certain days of the week, making people more likely to go see movies on those days. Movie revenues increase over time, presumably due in part to inflation. After redoing the regression while also incorporating the length of a movie title the MSE was lowered slightly to 1.5e+16 and the R^2 went up to a 0.647.

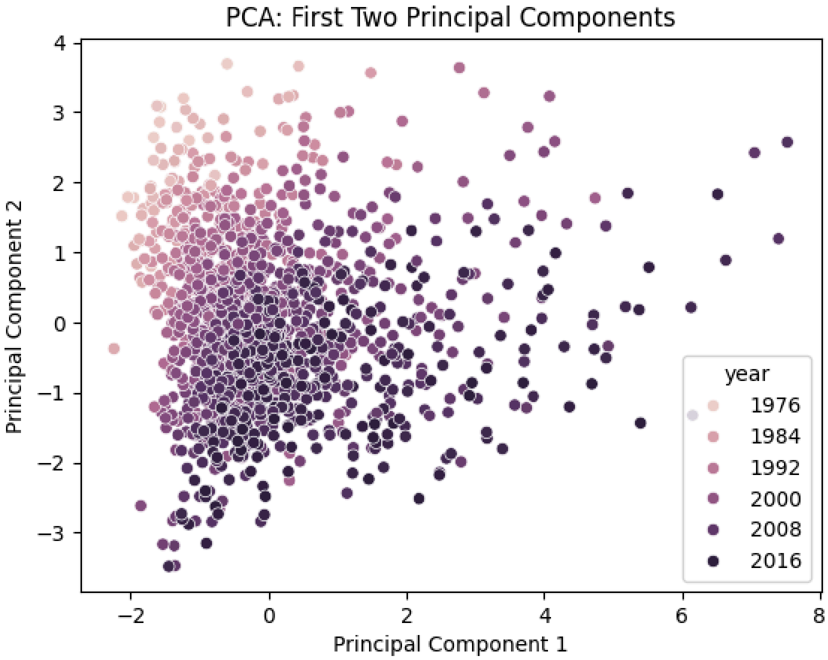
We then ran a second regression with fewer variables in hopes of reducing the overfitting of the model. As a result the R^2 was improved to a 0.66 and the MSE was lowered to 1.43e+16, possibly due to removing year so likely reduced noise or multicollinearity as could have been correlated to other features like budget, as budgets probably increased over time due to inflation.



As shown in the figure above there are more points in more recent years. This could be because the dataset mostly focused on recent movies (or there was just more easily accessible information for recent movies) or it could be that in recent years more movies come out each year. An interesting area for future study might be comparing the rate of inflation to the increase in revenue over the year to see if the revenue of movies increased proportionally to inflation or if movie quantity had a negative impact on quality (as determined by revenue).



It is also interesting how the vote average doesn't directly correspond to higher revenue with the highest earning movies seeming to have around 6 or 7 average opposed to 8. This could mean that large blockbusters with super high budgets and famous actors, like *Transformers*, are not the most well received by critics but are simply meant to get people through the door.



PCA allowed us to reduce the dimensionality of our data set and standardized how features contribute. Principle component 1 helped capture the largest variance in the data set by combining features that were likely to be highly influential like budget and vote count. Principle component 2 captures the second largest proportion of variance and is used to help include patterns not explained by PC1. The results of the PCA showed a moderately strong prediction with a R-squared value of 0.643 on the test set. The PCA graph shows how more recent movies, which were represented by the darker colored dots, could suggest trends in increasing budgets and higher audience interaction reflected by higher vote counts in more recent years. The more lightly colored points representing older movies were more grouped on the upper part of the y-axis and left of the x-axis showing that older movies associated with higher PC2 values and PC2 were able to capture some characteristics of older movies better than PC1. This could likely be due to less audience votes and a more even budget across all movies of the time. Overall the use of PCA allowed us to highlight differences between older and newer movies through simplifying the model which provides meaningful insights in patterns and trends in the data as changes in the movie industry impact revenue over time.

When we analyze the performance of the variables we discover a few things. Budget is strongly correlated with revenue (+91.2M coefficient), rather expected, as larger budgets typically translate to higher production quality and marketing reach. Vote count was shown to have a high correlation with revenue (+109.4M coefficient), suggesting audience engagement as a strong predictor. Also very understandable since movies that generate strong opinions, especially positive, would likely inspire more people to see a particular movie. The release month shows positive effects for December (+34.5M) and November (+36M), which align with holiday seasons when blockbuster releases are common. As for the gender of the director, male directors show a slightly negative coefficient (-2.2M), with missing gender values (NaN) also underperforming (-1.3M). These differences are minor and may not be statistically significant. Finally the overall model performance, R^2 Score (0.647 and 0.66 for the reworked model), indicates a moderately strong model. There's room to improve by addressing noise and adding meaningful predictors. MSE (1.5e+16): Likely due to the large range of revenues (e.g., small indies vs. blockbusters like Avatar).

Our analysis reveals that several production and release factors, particularly budget, vote count, and release timing, significantly impact a movie's financial success. Budget was the most influential predictor, which aligns with the understanding that higher investments in production and marketing generally yield higher revenues. Audience reception, represented by vote count and average ratings, also demonstrated a positive correlation with revenue. Additionally, release timing, especially in November and December, showed a strong seasonal influence, likely due to holiday-driven audience turnout.

The predictive model achieved a reasonable R-squared score of 0.647, suggesting a fair ability to explain variability in movie revenues. However, high MSE values indicate potential challenges in capturing the extreme variations between blockbuster hits and less successful films. Incorporating more sophisticated techniques like PCA or LASSO regression could further refine the model by reducing multicollinearity and overfitting, improving generalizability.

Future directions could explore inflation-adjusted revenue trends, deeper demographic analysis of directors and audiences, and the impact of title characteristics on success. These insights can provide more nuanced recommendations for the film industry, aiding in optimizing production strategies and release planning.

In conclusion, our analysis reveals that several production and release factors, particularly budget, vote count, and release timing, significantly impact a movie's financial success. Budget was the most influential predictor, which aligns with the understanding that higher investments in production and marketing generally yield higher revenues. Audience reception, represented by vote count and average ratings, also demonstrated a positive correlation with revenue, showing that films with broad appeal and receiving critical acclaim tend to do better overall at the box office. Additionally, release timing, especially in November and December, showed a strong seasonal influence, likely due to holiday-driven audience turnout.

Key Findings:

1. Budget as a predictor:

   The correlation between budget and revenue seen in our results (+91.2M coefficient) shows that larger investments in movie production and marketing lead to better success and these findings align with industry practices where large-scale films from well known studios who can spend a lot of money do the best at the box office.

2. Audience Reception:

   Higher vote count positively impacted revenue (+109.4M coefficient), reflecting how films that created buzz and led to audience engagement had increased revenue. However, vote averages did not always mean greater revenues. This may suggest that blockbusters that are aiming to make the most money may prioritize appealing to the greatest range of audiences while not caring as much about the opinions of critics.

3. Seasonal Trends:

   The impact of releases in November (+36M coefficient) and December (+34.5M) shows the importance of studios releasing movies around the holidays. Studios may benefit from scheduling releases during this period to take advantage of increased audience numbers.

Model Performance:

Our regression model was able to achieve a reasonable R-squared score of 0.647, and by removing some variables, was able to get a R-squared of 0.663, indicating a moderate amount of explanatory power. Our high MSE of 1.502 shows the challenge of modeling such a wide range of revenue outcomes across many different studios including indie films and massive blockbusters.

Future Directions:

To build upon our findings, future research could explore the following extensions

1. Incorporating Advanced Models:

   Leveraging machine learning algorithms like PCA, LASSO, or neural networks could help reduce the multicollinearity. It would better help the capture of nonlinear relationships between predictors and revenue.

2. Audience Demographics

Investigating how certain audience demographics like age, gender, location, and income could lead to a better understanding of predictors into film success.

3. Movie characteristics

Looking into how certain movie genres or subject matter lead to higher revenue and audience reception could help optimize marketing strategies and give studios insight into what type of movies will perform the best.

4. Inflation Adjusted Trends

Standardizing the budget and revenue for inflation and comparing trends over time could lead to a better understanding and insight into long-term shifts in box office performance.

Final Thoughts:

This analysis provides a good foundation for how certain variables affect movie revenue and provides insights for how studios can improve performance. By addressing certain limitations, implementing more specific data, and using advanced algorithms, further research can improve understanding and create more accurate models. As the film industry evolves, a need to adapt to compete against streaming platforms is highlighted even more. Integrating such models will remain crucial for a competitive advantage and staying relevant in a highly competitive landscape.

Bibliography

1. Ck, N. (2024, September 30). *Movies and directors: Dataset for film analytics*. Kaggle. https://www.kaggle.com/datasets/nayanack/movies-and-directors-dataset-for-film-analytics