DS3000 Project Final Report

BlockBuster Generator

Molly Varrenti
Ben Weiss
Sheena Kaw

Foundations of Data Science
Professor: Xiaoyi Yang
April 21, 2024

INTRODUCTION

Examining a movie's success involves delving into a multitude of factors, ranging from its cast and crew to its genre and more. Our investigation focuses on analyzing these influential elements using a dedicated database. We will delve deep into the highest-rated films from past decades to unravel the fundamental drivers behind their acclaim. This exploration will provide valuable insights for predicting potential blockbuster hits in the present year. Understanding audience preferences holds immense importance for filmmakers striving to create successful cinematic endeavors. To conduct our research, we will utilize web scraping techniques to collect data from the IMDB database (Chatzopoulou et al., 2010). The quest to forecast a movie's success prior to its release has intrigued scholars since the early 1980s. Numerous researchers have sought to identify the factors impacting a film's box office performance. Predicting box office revenue (BOR) has become increasingly crucial in the film industry due to the involvement of stakeholders like investors, advertisers, and cinemas (Wang et al., 2020). By extracting data from the IMDB database, we can assemble a comprehensive dataset comprising movie ratings, genres, directors, and actors. Scrutinizing this dataset can furnish invaluable insights into the elements driving a movie's triumph, thereby assisting producers in strategic decision-making, such as casting renowned actors and directors.

DATA SOURCE

From the list of movies on IMDB, we selected about 1,700 of them released between 1970 and 2024. There are far more on the website, but this would require a great amount of computational power, and we were concerned that including a number of small or irrelevant movies would not be beneficial to our goal of trying to determine what the top movies have in common. Scraping this website for names, worldwide gross, directors, lead actors, genre, and more allowed us to accurately analyze our research question. We were primarily trying to scrape the worldwide gross for each movie, and any further data we could acquire would just be a bonus as we try to understand the trends that make movies perform well at the box office.

There was some level of cleaning that we performed after our initial rounds of scraping from the website. Most big movies had a complete set of fields, but some were missing a category or two, and this had the potential to throw off our whole technique during scraping. Certain categories were swapped or null, so we created a few Python functions to toss out illegitimate data points. Our final CSV contained only movies that had a worldwide gross and a sufficient number of other categories filled out.

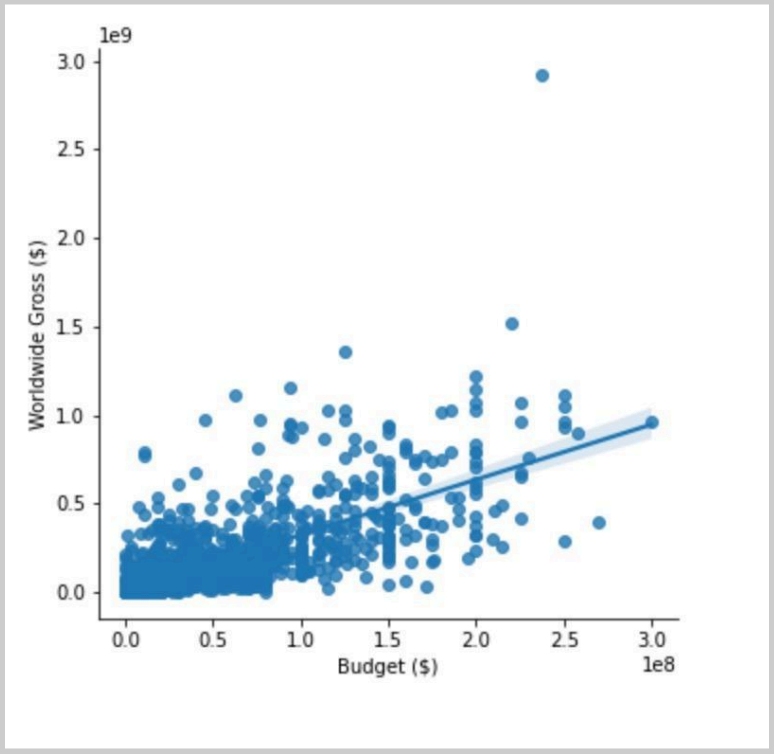
SOURCES

- Chatzopoulou, G., Sheng, C., & Faloutsos, M. (2010, March 1). A First Step Towards Understanding Popularity in YouTube. https://doi.org/10.1109/infcomw.2010.5466701

- Wang, Z., Zhang, J., Ji, S., Meng, C., & Zheng, Y. (2020, August 1). Predicting and ranking box office revenue of movies based on big data. Information Fusion, 60, 25-40. https://doi.org/10.1016/j.inffus.2020.02.002

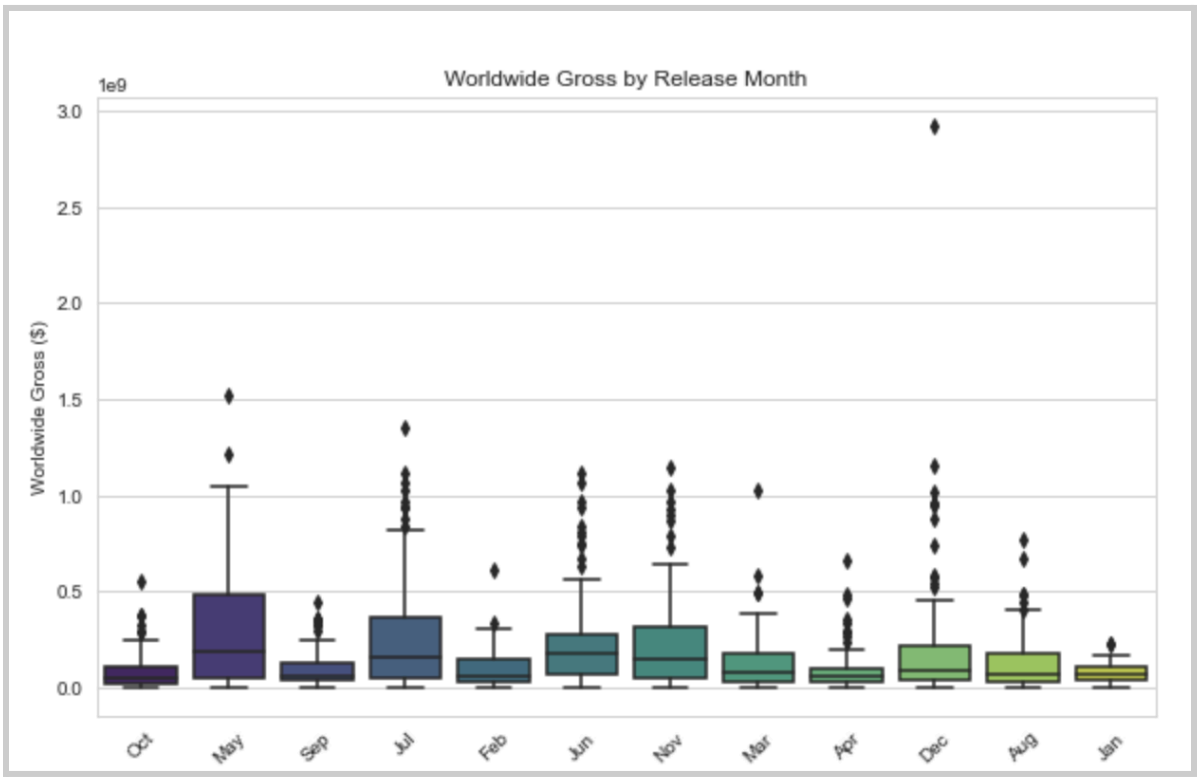
DATA VISUALIZATIONS

1. *Relationship Between Budget and Worldwide Gross:*



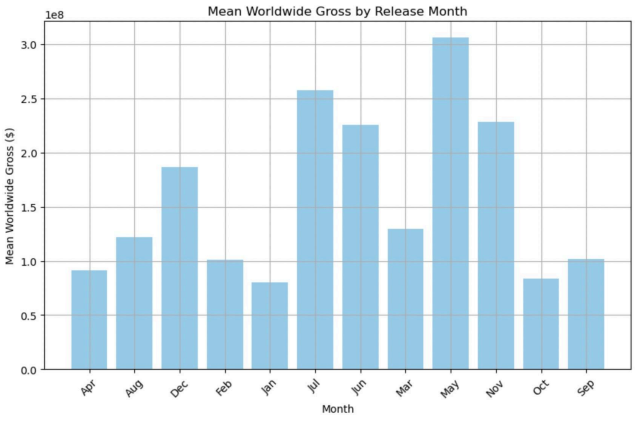
The most impactful analysis we did was plotting the correlation between a movie's budget and worldwide gross. As can be seen above, there is a strong relationship between budget and worldwide gross. Further analysis during our project went on to show that a budget is by far the most influential factor in a movie's success.

2. *Impact of Release Month of Worldwide Gross:*



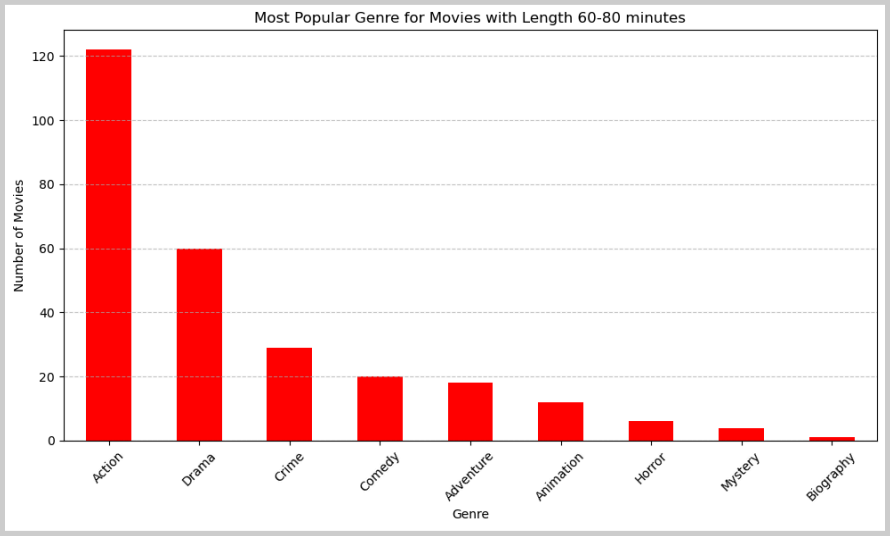
There was not a huge variation in worldwide gross across the board when comparing the 12 months, but it did help us identify a few peak times of year. The holiday season of November/December and the summer months emerged as the best times of year to release a movie. The month of May slightly edges over the other 11 months, with the highest mean and median worldwide gross over the data we analyzed, making it the most ideal for a movie release.

Corresponding check plot:



This check plot assesses a linear regression model's performance in predicting movie worldwide gross earnings. Each point represents a movie, with its position showing actual versus predicted earnings. The red dashed line signifies perfect prediction. Close clustering around this line indicates accurate predictions, while deviations imply inaccuracies. The spread of points reflects prediction variability. This plot visually evaluates the model's performance, indicating areas for improvement.

3. *Most Popular Genre for Our Decided Length:*



We determined in a separate visualization that 60-80 minutes would be the best length for a blockbuster movie, and this simple bar graph shows which movies tend to be in this runtime range. This showed that the vast majority of short movies of 80 minutes or less are in the action genre, and this was part of our rationale for deciding that our movie should be an action film.

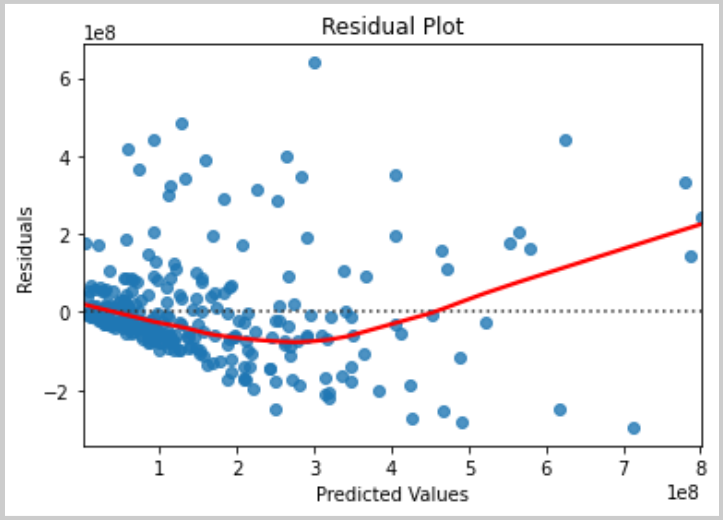
STATISTICAL MODELS

Model 1: Linear Regression

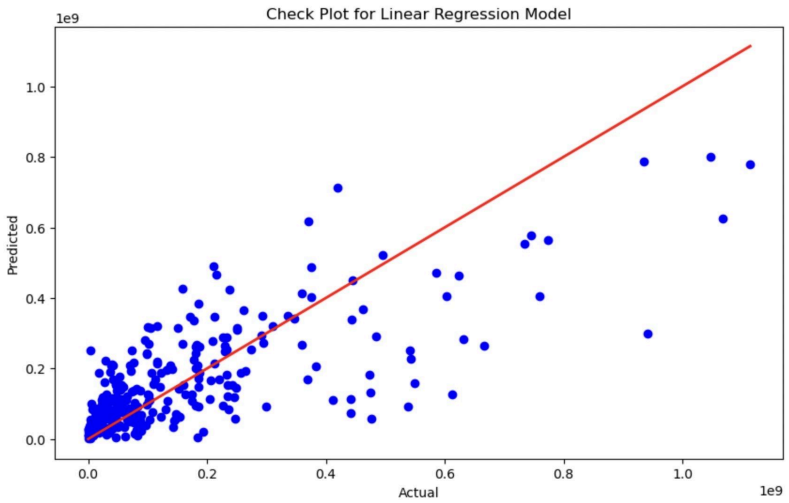
Our first model was a linear regression model with world wide gross as a dependent variable and movie budget as the independent variable. We chose length and budget as the independent variables as they are the two numeric variables and would be a practical choice to include in the model. Linear regression models provide coefficients that directly represent the relationship between the independent variables (features) and the dependent variable (target). In this case, the coefficients indicate how changes in the budget and length of movies relate to changes in the worldwide gross. Linear regression is a simple and easy-to-understand modeling approach, which makes it suitable for initial exploratory analysis and as a baseline model.

The results of this model yield a coefficient for budget to be approximately 3.09 which indicates that for every unit increase in budget, world wide gross is expected to increase by about 3.09 million dollars. Similarly, for the 'Length (minutes)' feature, the coefficient is approximately −767,487, suggesting that as the length of the movie increases by one minute, the predicted worldwide gross decreases by around $767,487. To continue, the MSE value is quite large, which indicates that the models predictions are far from the actual values on average. The r squared value represents the proportion of variance in world wide gross that is predictable from budget and length. The value from our model is 0.55, indicating that about 55% of the variance can be explained by our independent variables with our model. While this indicates some level of predictability, there is still a substantial amount of variance that is not accounted for.

Furthermore, check plots were created for this model. The first plot generated is a residual plot and is shown below:



Corresponding Check Plot:

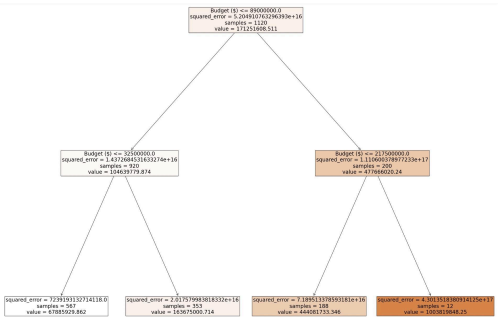


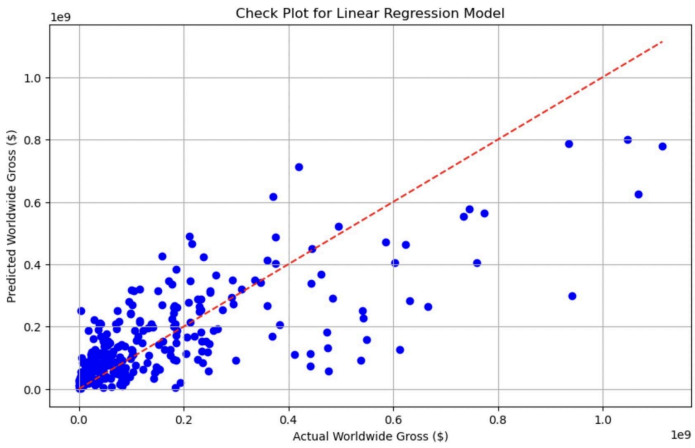
The check plot compares the model's predicted values with the actual values. Each point represents a data instance, with the x-coordinate as the actual value and the y-coordinate as the predicted value. The diagonal line represents perfect predictions, while deviations indicate prediction errors. Metrics like Mean Squared Error and R-squared offer quantitative assessment. Ultimately, the plot visually evaluates the model's accuracy in predicting the target variable.

Model 2: Decision Tree

The second model that was generated is a decision tree model. This was chosen as decision trees can help to capture nonlinear relationships between features and the target variable. Since the data is complex and our linear regression model was unable to adequately capture relationships, a decision tree model was generated next. Decision trees are easily interpretable, so they offer a clear understanding of how the model arrives at its predictions. This is valuable especially to our data because stakeholders may want to understand the reasoning behind the predictions of budget and length.

After creating this decision tree, we calculated the feature importances to determine which of the two features has a larger effect on the model.The results of the feature importances are 1.0 for budget and 0.0 for length. This indicates that length contribution is insignificant compared to budget in the model. The MSE value for this model was approx 2.11 and the r squared value is around 0.46. The MSE value is quite high, indicating that the predictions are not very accurate or that there is high variance in the data. The r squared value then, suggests that about 46% of the variance of the target variable can be explained by the model. This means that model is not capturing all the factors contributing to the prediction.





This check plot shows the mean worldwide gross earnings for movies released each month. Each bar represents the average earnings of movies released in a particular month. By comparing the heights of the bars, you can quickly assess which months tend to have higher or lower mean earnings. For instance, if you observe taller bars for certain months, it indicates that movies released during those months tend to have higher average worldwide gross earnings. Conversely, shorter bars suggest lower average earnings for movies released in those months. This visualization helps in understanding the variation in movie earnings across different release months and can provide insights into the seasonality or trends in movie performance throughout the year.

CONCLUSION

The goal of our project was to create the perfect movie to dominate the worldwide box office for this upcoming year. We sought to find the ideal budget, director, lead actor, time of year, and so on to create perfect conditions based on historical trends. During our analysis of over 1,700 past movies, we believe that we found a few key ideas for us to utilize.

BUDGET

The most common trend in our research was the relevance of a movie's budget. A visualization showing the relationship between movie budget and worldwide gross is almost completely linear upwards. That is to say, historical trends show that a higher production budget almost always correlates to higher worldwide gross. It seems rather obvious to say that we want as high of a budget as possible for our movie, but our analysis shows that this is the most important and determining factor on our upcoming film being a blockbuster. The 5 most expensive movies we saw had a budget of around $250 million, and this would be the best budget for our movie to have a high chance of topping the charts this year.

RELEASE DATE

There are a few major times of year when major movies seem to be released. Thanksgiving, Christmas, Labor Day, and parts of the summer always seem to be the most littered with the year's biggest movies. Our analysis shows that there is a good reason for this. Like us, major production companies have done research on box office trends and found a few peak times of year.

There were a few times of year that were better than others, but overall there were several release dates throughout the year that had comparable rates of success. Each month also seemed to have at least one or two heavy outliers, implying that a good movie has the potential to succeed regardless of the time of year.

Even though there are several good release dates and it was not as big of a factor as we originally believed, we have decided to release our upcoming movie during Memorial Day weekend. The month of May had the highest mean and median worldwide gross across all of the months, and we believe that Memorial Day weekend would be the best time of the month to maximize viewership. The summer months had some of the best worldwide gross, and releasing our movie over Memorial Day weekend, which is largely considered to be the kickoff of summer, would give our movie a great chance to be a top performer.

LENGTH

Another factor that we took into consideration was the approximate length that our movie should be. There was not a large distinction to be made between movies of varying lengths, but around the 80 minute mark the mean worldwide gross started to take a slight dip. For this reason, we have made the decision to keep our movie relatively short, ideally less than an hour and a half.

GENRE

Our analysis did not find a strong correlation between worldwide gross and several of the major genres, but this gap did start to widen when taking the budget, length, and release date into consideration. Taking the 3 factors from above into consideration, our analysis showed that an action movie would be in the best position to perform well.

The majority of movies that have a runtime of 80 minutes or less are action movies, and these are also the most prevalent and popular during the summer release season. Over 120 action movies in our dataset ran between 60 and 80 minutes, which is over twice the amount of any other genre.

SUMMARY

There are a few other considerations for a movie production, such as actors, directors, production company, film type, and so much more, but we believe that we have identified several key principles to have the year's next big blockbuster.

We also believe that there are a few things that we could have done better. While having 1,700 movies was plentiful, having more movies, particularly more movies from the last 5-10 years, could have given us more robust information about our decisions. We quickly identified that budget and release date were important factors, so from there we could have started to narrow down our analysis. Instead of constantly doing analysis of everything, we could have focused down to our specific criteria. We did this for deciding on our genre, but it could have been possible to do this even further and for more factors.

After analysis of over 50 years of movies, we have determined that our blockbuster will be a short action film coming Memorial Day 2025.