Predicting Film Financial and Critical Success using Graphical Analysis and Machine Learning Techniques

By Jacob Lush and Jacob Strozyk

Table of Contents

Executive Summary

Which genres of films generate the most revenue?

Analysis from our 2017 data reveals that Animation and Family genre films generate the most revenue, and TV Movie genre films produce the greatest revenue to budget ratio.

Which genres of films have the best rating?

Analysis from our 2017 data shows that War films have the highest average ratings, followed by Music and Thriller films.

Is there a correlation between film rating and film revenue?

Our analysis shows that film ratings and film revenue had a positive correlation; films which were received better critically generated more revenue. Films which were more highly rated also had a better revenue to budget ratio.

How accurately can we predict past ratings from the information gathered from past films using a regression machine learning model?

Using our mode, we were able to rate past films with an average error of about 0.5 to 0.7 (on a rating scale of 1 to 5). We think this is a pretty accurate model. If we were to introduce more dimensions to our model (e.g. adding director, producer, etc. information for our model), then we believe that we could improve accuracy.

How accurately can we predict future ratings from the information gathered from past films using a regression machine learning model?

Using our model, we were able to rate future films with an average error of about 0.7 to 0.8 (on a rating scale of 1 to 5). We believe this to be a good level of accuracy. If we were to introduce more dimensions to our model (e.g. adding director, producer, etc. information for our model), then we believe that we could improve accuracy even moreso.

How accurately can we predict past revenue from other information gathered from past films using a regression machine learning model?

Using our model, we got huge errors. We have come to the conclusion that due to the massive range the revenue can have mixed with the very large sample size, there is just too much variation to have an accurate model.

Motivation

Being able to predict the ratings for a film would be useful for finding films that may be of interest that are not yet released. For example, if we know a lot about an unreleased film's details, then we can get an idea of what the critical reception would be before it even comes out. This could have personal uses such as deciding which films to watch on release date, but also could be used by film company executives to get an idea of which film proposals they receive would be most successful.

Understanding which genres of films receive the best ratings and which make the most revenue would also be helpful for making informed decisions about what media we are to consume when combined with knowing how ratings and revenue correlate. For example, if you want to decide which film to watch when you can only watch a single one, being able to make a decision based on previous analysis of film statistics would be helpful in making your decision. Of course, film companies would also love to have this information.

The analysis in this report helps both individuals who consume film media and also companies which produce said media make more informed decisions. Is it better to watch a War film which you know made hundreds of millions world wide, or would it be better to watch a Comedy film which you already know has a 4 star rating? Would it be more profitable to make a big budget musical or a small budget TV Movie? These are questions which can be guided toward an answer with our analysis.

Datasets

For our analysis, we used two datasets.

The first dataset is “The Movies Dataset” created by Rounak Banik. This dataset contains information gathered from over 45,000 films listed on MovieLens. The films in this dataset only go up to films released in 2017. We used this dataset for our graphical analysis and for training our machine learning models. We refer to this as the “2017 dataset.”

“The Movies Dataset” is available here:
https://www.kaggle.com/rounakbanik/the-movies-dataset

The second dataset is the “IMDB movies extensive dataset” created by Stefano Leone. It contains information for over 85,000 films that are listed on IMDB. It contains information on films that were released up to 2020. We used this dataset to find out how accurate our machine learning models from the previous dataset were. We refer to this as the “2020 dataset.”

The “IMDB movies extensive dataset” is available here:
https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset

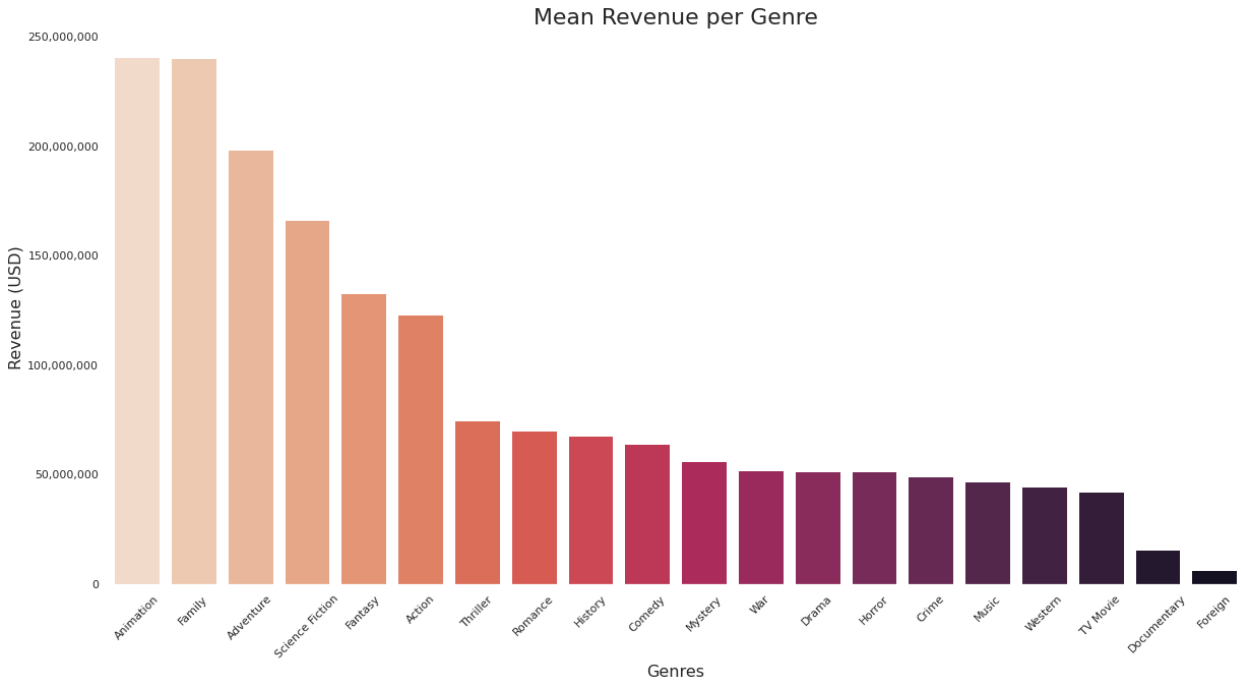
Methodology

1. Parse data from the 2017 and 2020 datasets into pandas dataframes. Make any necessary modifications to the data for later steps. Use the pandas library.
2. Use data visualization techniques to display genres, their respective mean revenue and revenue to budget ratios. Use the seaborn and matplotlib libraries.
3. Use data visualization techniques to display film ratings, their respective mean revenue and revenue to budget ratios. Use the seaborn and matplotlib libraries.
4. Analyze variables such as budget, genre, and rating using machine learning techniques in order to accurately predict revenue for any given film in any given genre. Use the pandas and scikit-learn libraries.
5. Similar to question 4, but switching the rating feature with revenue.

Results

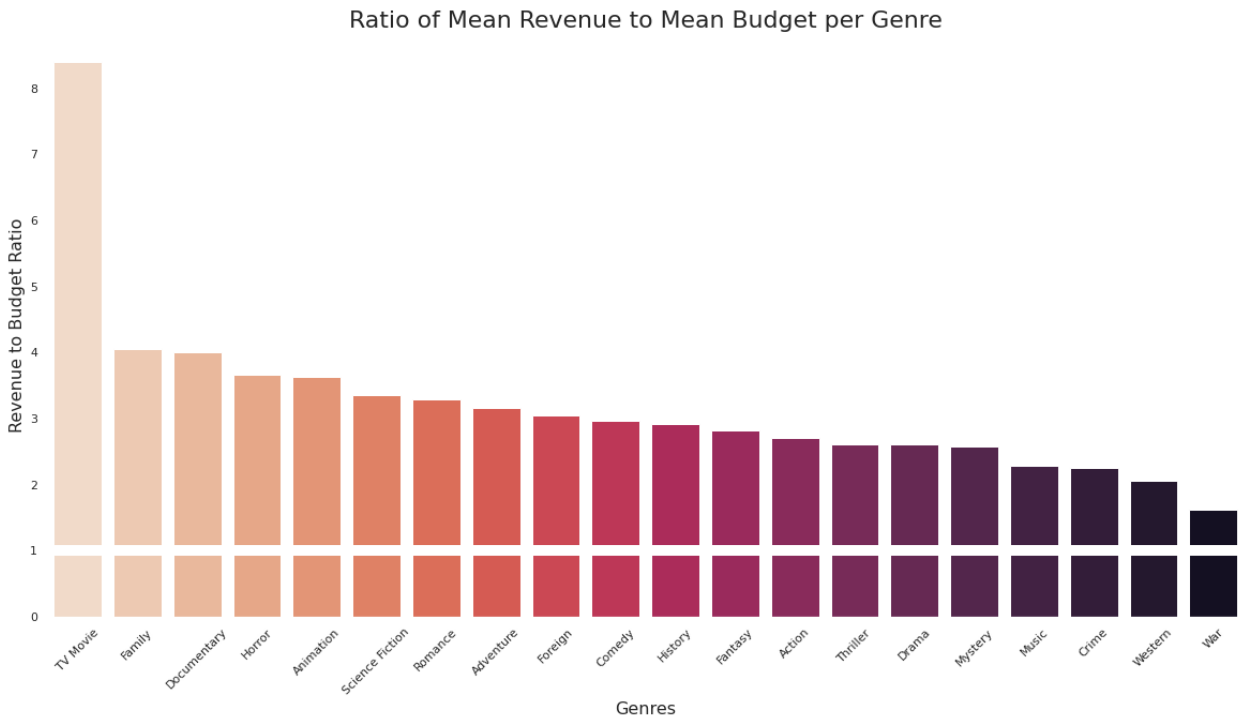
Which genres generate the most revenue?

To find out which genres generate the most revenue, we looked at the 2017 dataset. We used the pandas, seaborn, and matplotlib libraries to read the dataset, modify its data in the ways we needed, and then graph the result. Here are the graphs that we produced to answer this research question :



The above graph shows the raw average revenue generated by each genre. It is not surprising that animation films which usually have huge budgets and family films which have a wide audience generate the most revenue on average. Foreign and documentary films generate surprisingly low revenue on average. This graph is good at showing how “big” a genre is on average. However, it does not give a good idea of which films would actually be the most profitable. This is due to the difference in budgets. An indie foreign film will usually not have the same budget as a hollywood blockbuster, so it is unfair to compare the raw amount of revenue generated when trying to find what type of film would be most profitable*. The next graph compares the ratio of revenue to budget for films. A value of 5 would mean that a genre of film, on average, generates 5 times its budget as revenue (for a profit of about 4 times the budget).

* revenue is the total amount of money made, profit is the revenue with expenses taken out.

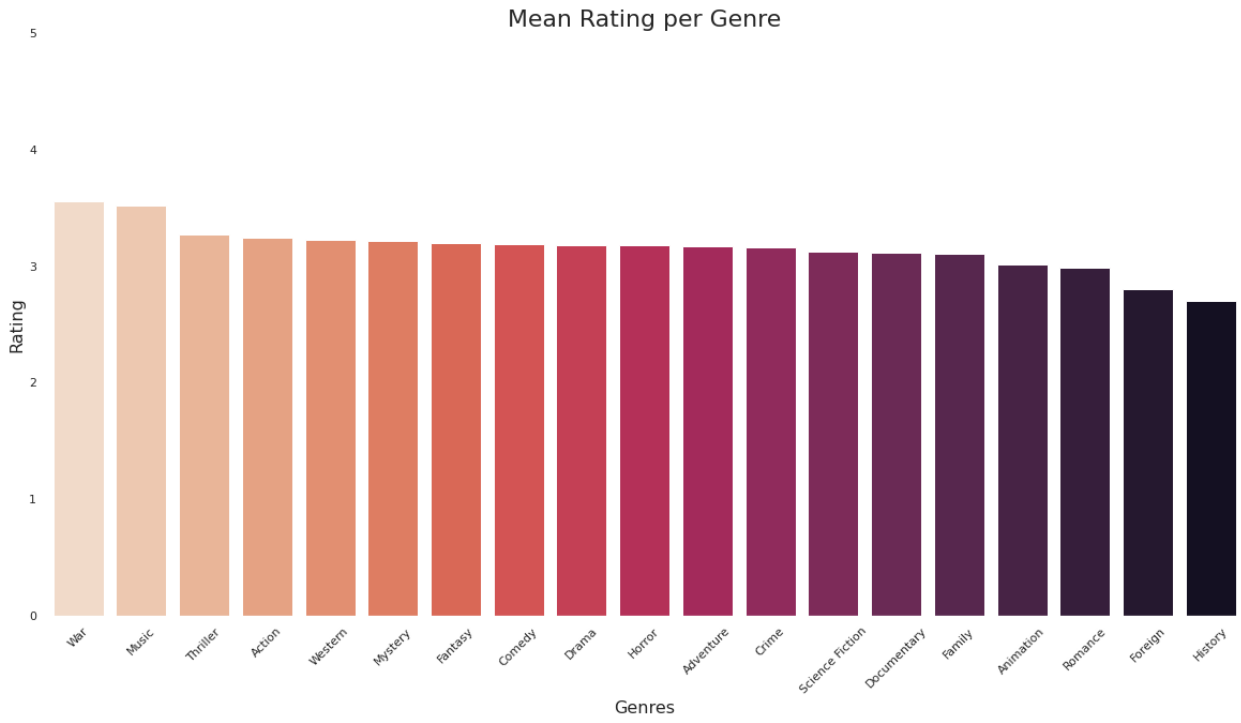


This graph gives a better idea of what types of films create the greatest profit for investment. The white line on the graph represents the point at which the revenue of a film equals its budget (i.e. breaks even). TV Movies are an outlier; they generate a huge amount of revenue compared to their budget. Animation films may have generated the most raw revenue, but their place in this graph shows that while they are great at generating profit, they are not as dominant as they were in the previous graph. Foreign films rise greatly in this graph, showing that their small budget makes up for their small revenue generation.

We conclude that the genres which produce the greatest amount of revenue are animation and family films. However, TV Movies generate the largest revenue relative to their budget. Choosing a different genre of film has a significant effect on how much revenue a film will generate or how much it will generate relative to its budget.

Which genres have the best rating?

In order to find out which genres had the best ratings, we performed the same steps as the previous research question to gather and prepare data. The graph below shows the mean rating per genre.

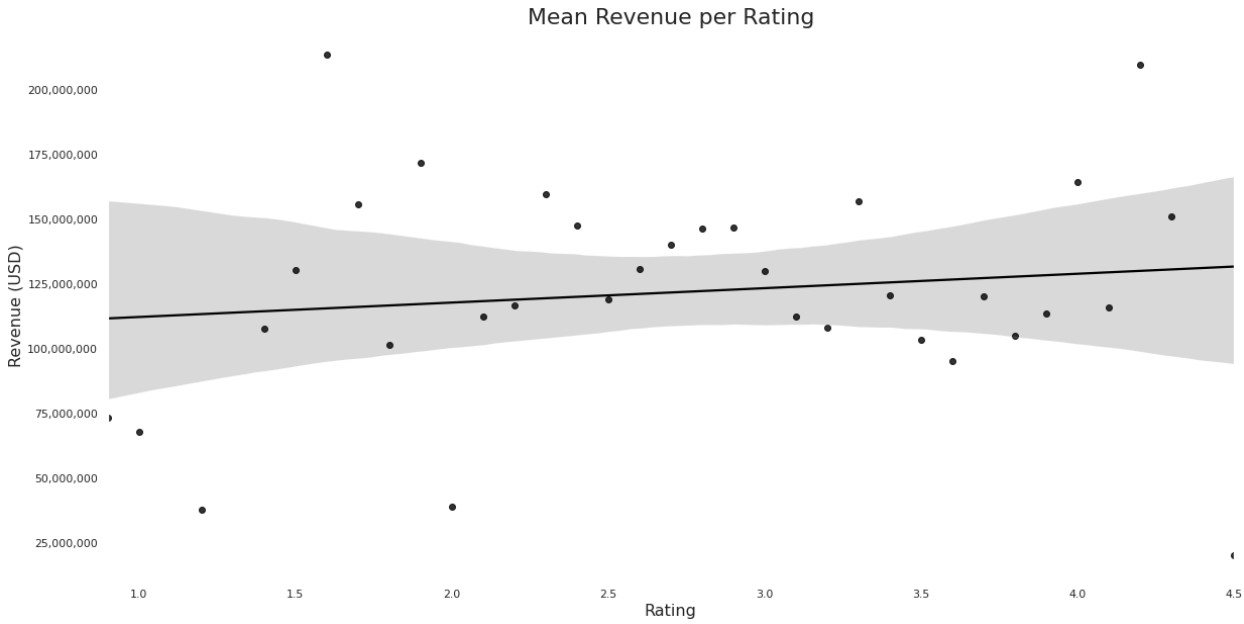


The graph shows that the difference in rating between genres is not dramatic. The difference in mean rating between the top and bottom genre is less than 1. The highest rated genre on average is the War genre. The lowest rated genre on average is the history genre.

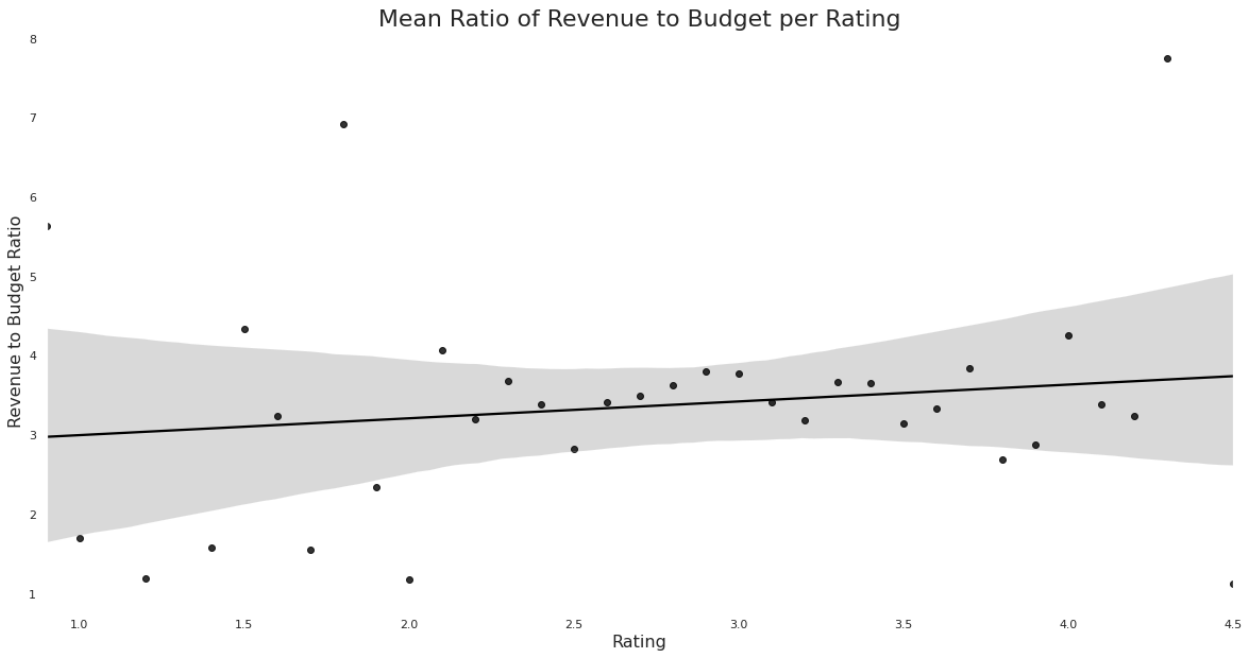
We conclude that genre is not critical in determining a film's rating. However, on average, War films have the highest ratings.

Is there a correlation between film rating and film revenue?

We repeat the same process as the last two questions in order to prepare our data. This time, we are not interested in genres. Instead of using a bar graph, we now use a linear regression graph. We group films by ratings down to the tenths place (e.g. 2.5, 2.6, 2.7). We can now compare the revenues for each rating.



The graph above shows that there is a positive correlation between rating and revenue. The slope of the line shows that films which have higher ratings on average generate more raw revenue. However, like our first research question, we may also be interested to see if these films are more profitable. We repeat the same process as before by dividing revenues by budget to find a ratio which gives a general idea of how profitable a film is.



This graph shows the ratio of revenue to budget for different film ratings. There is still a positive correlation! This means that films that are received better critically are also more profitable. Surprisingly, this graph looks almost identical to the previous graph. This is unlike the previous research question where the same process greatly altered the result.

We conclude that there is a positive correlation between film rating and film revenue. Not only this, but there is also a positive correlation between film rating and film profit. If you want to make more money, and make it more efficiently, then strive to create a highly rated film.

How accurately can we predict past and future ratings from the information gathered from past films using a regression machine learning model?

This was the most complex research question because it involved multiple steps and a lot of data manipulation. We used scikit learn's regression tree model and used data from our 2017 dataset to train it. We used features such as budget, revenue and genre and film rating for labels. Using the 2017 data alone generated a pretty accurate model. Using the 2017 dataset for test values, we found that our model had a mean squared error between about 0.5 and 0.7. This means that our model was off, on average, by about 0.5 to 0.7 for the rating. The last time we ran this code, we got an exact mean square error of 0.49149.

However, to see if this model could accurately predict films from the future, we would have to use our 2020 dataset. We trained our model in the same way mentioned previously with a few modifications to make sure that the model would be compatible with the 2020 dataset. This required, for example, renaming some genres, dropping a few specific genres, dropping foreign currencies from our profits in the 2020 data, and calculating revenue for the 2020 dataset. In

short, we had to make many changes in order for the 2017 model and the 2020 data to be compatible.

Once our model and data were compatible, we used the 2017 model to predict films using features from the 2020 data. Afterward, we compared the predicted values against the 2020 test values which we had generated. Unsurprisingly, the error for this was a little bit larger than the previous model. We found that the error ranged from about 0.7 to 0.8. The last time we ran the code, our error was 0.74067. What this means is that our model from 2017 was able to predict ratings for films that were released in the future (relative to the data) with relatively good accuracy!

We conclude that we can accurately predict future film ratings within about 0.7 to 0.8 rating. The most exciting part of this result is that it can be improved relatively easily. If we were to introduce more/different features for the machine learning model, we could increase the accuracy of the model. Currently, we only have three features being analyzed (budget, genre, revenue). If we were to introduce features such as runtime or information about who made the film, then we could improve the model.

How accurately can we predict past revenue from other information gathered from past films using a regression machine learning model?

This model initially drew some concern. We used scikit-learn to generate a model using our 2017 dataset, setting aside the ‘revenue’ as the label and ‘rating’, ‘budget’, and ‘genre’ as the features. The concerning thing was that our error was astronomically high! We spent a few hours trying to figure out where we went wrong in our implementation, but have come to the conclusion that since revenue is so volatile and can range widely from film to film, we are bound to get a large error due to the large sample size and range of values the revenue can have.

Challenge Goals

Our completed challenge goals were:

- Machine Learning
- Multiple Datasets

We decided to not pursue result validity because it would require a lot of research, math, and time to implement. We believe that our calculated errors are satisfactory to ensure that our results are reasonable.

Work Plan Evaluation

Our original work plan played out pretty well. Although, our time estimates for some of the steps were inaccurate. Some things took way longer than expected, while others were completed more quickly. Specifically, using the 2017 machine learning model to predict future movie ratings was a bigger challenge than expected, as it involved a lot of data cleaning due to the formatting differences between the datasets.

We also decided to drop the research question: “How do original films compare to films from franchises in terms of ratings and revenue” due to the difficulty of selecting film sequels from the dataset (as they were formatted in a variety of ways), and the limited time frame we had to get everything else completed.

Testing

The majority of our testing was done by using manual inspection. To test our dataset transformations, we would use print statements to view the data before and after the modifications. For our graphs, we would use printed values from our datasets to ensure the values our graphs were outputting seemed reasonable. For our machine learning models, we would print out the mean squared error using the testing method we learned and used in our homework and lessons, as well as print statements when working with lists and dataframes.

Collaboration

To display revenue using commas on our graphs, “6,000,000” instead of “6000000” for example, we used code from user “falsetru” on StackOverFlow. The link to the thread with the code can be found here:
https://stackoverflow.com/questions/25973581/how-do-i-format-axis-number-format-to-thousands-with-a-comma-in-matplotlib

To filter our genres, we used the isin function. We learned about this implementation from the user “atinjaki” on StackOverFlow. The link to the thread with the code can be found here:
https://stackoverflow.com/questions/59275119/how-can-i-filter-single-column-in-a-dataframe-on-multiple-values

All other code in our project is original and based on material from lectures, checkpoints and sections from the CSE 163 Intermediate Data Programming course.