Analyzing Box Office Revenues and Profits

Zane Brown

[Zjbrown152@gmail.com](mailto:Zjbrown152@gmail.com)

4/25/2023

**Executive Summary:**

In the analysis of movie revenue, I examined the performance of different models, including Random Forest, Logistic Regression, and Multiple Linear Regression. I decided that I was also going to create a profit classifier variable that represented if a movie made a profit higher than the median profit. I did this due to a high degree of skewness in revenues in my dataset. Among these models, Logistic Regression emerged as the most effective predictor of movie revenue. However, it is important to note that our models were trained solely on historical data due to the unavailability of live data, which may limit their real-time accuracy.

The analysis uncovered strong positive correlations between vote count and revenue, as well as between budget and revenue. These findings indicate that both audience reception and awareness, as measured by the vote count, and financial investment in the form of budget significantly influence a movie's revenue. It suggests that movies with higher audience engagement/awareness and greater financial resources tend to perform better financially. Based on these findings, I conclude that budget plays a vital role in determining movie revenue. Allocating a higher budget to a movie increases its chances of generating higher revenue. Additionally, the positive correlation between vote count and revenue highlights the importance of audience engagement in driving financial success for movies.

To improve the accuracy of future analyses and predictions, I recommend the following actions. Firstly, gathering more data by including a larger sample of movies across different genres, release dates, and production budgets would improve the predictive capabilities of our models. A larger dataset would provide a comprehensive understanding of the various factors influencing movie revenue. Introducing additional variables would enable a more in-depth analysis. Variables such as marketing cost, social media presence, critical reviews, and competition within the same release window could offer valuable insights into other significant aspects impacting movie revenue. Lastly, seeking live data would be beneficial for creating more up-to-date models. Access to real-time data would allow us to incorporate the latest information, resulting in more accurate predictions that reflect the dynamic nature of the film industry and current market trends.

**Introduction:**

This project has three main objectives. The first objective is to explore how the revenue generated by movies in different genres differ from each other. For example, a certain genre may generate more revenue than another during specific parts of the year. Secondly, I will create a model that can accurately predict what the revenue of a movie will be based on a variety of different independent variables. I will run many different models to determine which performs the best and makes the most sense. Lastly, I am to identify an independent variable or variables that could be significant in determining what a movie's revenue could be.

**Project Details:**

The project consists of three parts, data collection, data exploration and predictive analytics. Below is a more in-depth explanation of what I did to accomplish each part.

Data Collection:

The data collection portion consisted of a combination of Python programming, IMDb, and TMDb (The Movie Database) API. Initially, I had access to a large text file containing over 12 million movie IDs sourced from IMDb. IMDb is the most used API for movie data researchers so naturally I wanted to utilize it. However, I encountered limitations due to the requirement of a paid subscription to access the IMDb API directly. Due to this, I turned to the TMDb API, which provided access to a range of movie data on popular/trending films and also pull specific movie’s data using an id search.

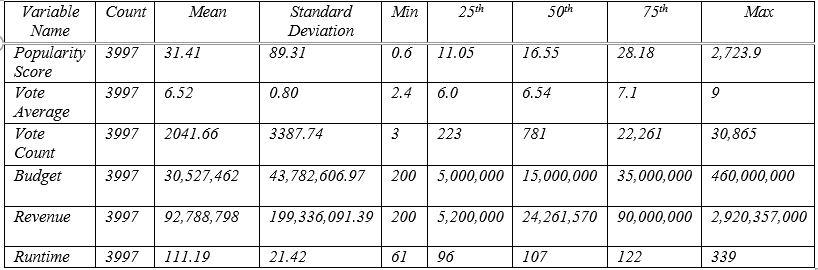
To extract and organize the movie data efficiently, I utilized the Python programming language, specifically leveraging the pandas library. With pandas, I could easily manipulate and structure the data into a coherent format. Additionally, Python allowed me to handle the data cleaning process effectively, eliminating movies with null values and ensuring data integrity. Upon exploring TMDb further, I found that a significant number of movies extracted from the TMDb API contained null values, likely due to their new or upcoming status. So I needed a way to supplement this data, this is where the IMDb IDs proved invaluable. By running the IMDb IDs through the TMDb API, I could retrieve additional movie information, effectively expanding the dataset's scope.

To process the movie IDs from the IMDb database file, I employed a for loop within my Python code. This allowed me to iterate through a sample of the IMDb IDs and query the TMDb API for relevant movie details. However, due to the size of the dataset, even a limited sample of 20,000 proved time-consuming to process. During the loop iteration, it was important to identify and exclude movie IDs that caused errors or failed to retrieve any information from the TMDb API. By implementing appropriate error handling mechanisms, I could ensure that only valid IDs were considered, thus improving the reliability and quality of the final dataset.

Upon completing the data collection process, I successfully compiled a refined dataset comprising 3,997 movies. This number represents the subset of movies that retained complete and valuable information after filtering out null values and erroneous IDs. These 3,997 movies were meticulously curated and ready for subsequent analysis. Now that I had a complete dataset that I was happy with, I needed to transfer my data into a data store. To do this, I established a connection to an SQLite3 database using the Python sqlite3 package. This facilitated the seamless integration of the extracted information into a structured format. As a result, I could easily manipulate and query the data for further analysis.

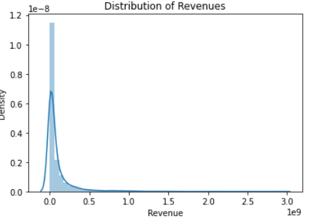
Data Exploration:

To perform my data exploration, I used Pythons statistics package to get summary statistics for my data set including mean, minimum, maximum, and standard deviation. You can see these for my dataset below.

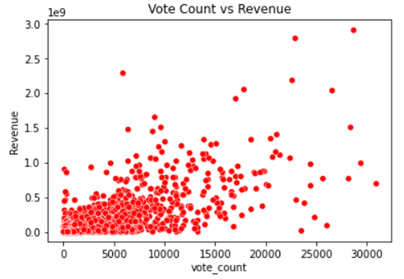
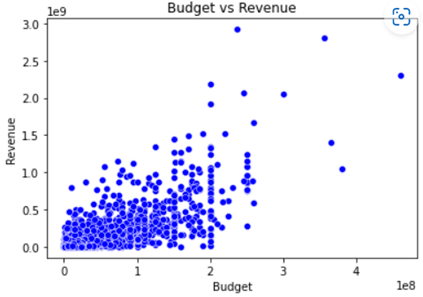


A couple things that stood out to me were the minimum values for budget and revenue. They both had a minimum value of 200. This caused me some concern because it is highly unlikely that a movie was made on a budget of $200 and likewise for a revenue of $200. This led me to further investigate to see if it was only one movie or a recurring issue. I found that this only occurred for 3 movies in the dataset and decided to just remove these movies from the dataset before I continued.

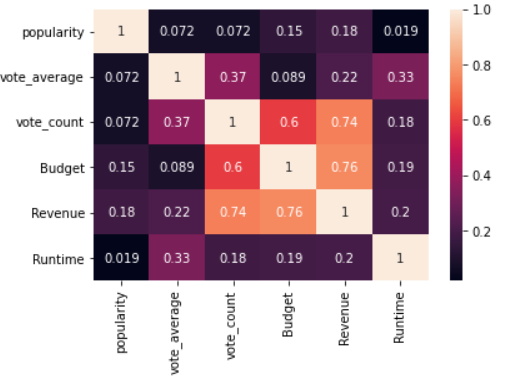
I also used python to create distribution plots and scatter plots to try to find any patterns or relationships present within the data as well as a correlation matrix. Most of the distribution plots of the variables in the dataset were highly skewed. Looking at the distribution plot of revenues, we can see below that while most of the data is in the lower end of the plot. However, movies that have an abnormally large revenue compared to the rest of the data causes the plot to have a significant right skew.



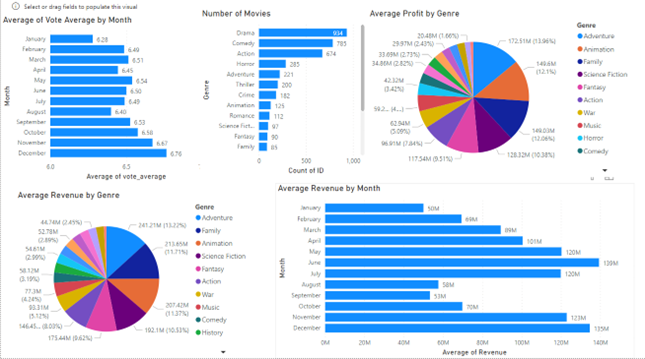
When making scatter plots I found two that seem to have a relationship present, budget/revenue and vote count/revenue. As you can see from the scatterplot of budget and revenue below, there is a generally positive relationship present between the two. The scatter plot for vote count and revenue looked very similar to this one as well. This indicates that generally for the dataset, as budget rises so will the revenue and the more votes a movie has the more revenue the movie generated.

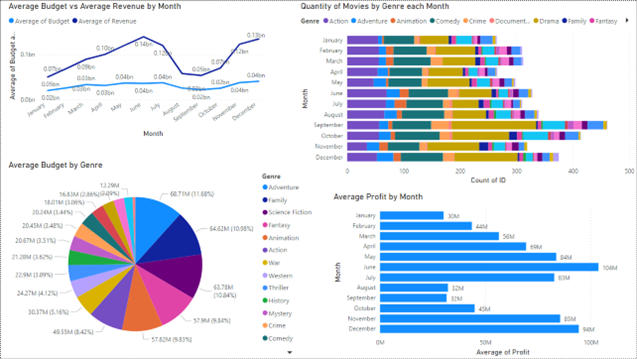


The correlation matrix below shows the correlation values present between the variables in the dataset. As you can see budget/revenue and vote count/revenue had the highest values. This further supports what I found in the scatter plots.



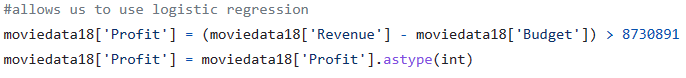
Lastly, I connected my data store, sqlite3, to Power BI to do further visual data exploration. Specifically to see trends across genres and different times of the year. This includes average revenue by genre, average budget by genre, average profits by genre and the number of movies in each genre released each month. Other visualizations that showed the average revenue by month and the average profit by month were included as well as the number of movies in each genre in the complete dataset.





Predictive Analytics:

I decided to create a predictive model for two different outcomes, one being the revenue of a movie and the other being a classifier for if the movie made a profit higher than the median. I chose to create a profit classifier because I felt that it would give more useful results in comparison to revenue due to the high skewness present in my dataset. To do this I used the median of revenues to run the following code that created the classifier column.



This allowed me to run logistic regression, decision tree, and random forest classification models on the data set for the profit classifier. For the revenue, I did multiple linear regression and random forest regression. However, before I could do any of these, I had to create dummy variables for my categorical variables such as genre and release date. I used just the month of the release date to create these dummy variables. I also found that I had to use a standard scaler to transform my data before fitting it to a model for all of the models except for multiple linear regression.

**Reflection:**

Throughout the course of my project, there were several positive aspects that contributed to its success. Firstly, one of the major strengths was the availability of a mostly clean and easily understandable dataset. The Movie Database API was easy to learn how to use, this allowed me to dive into the analysis phase without having to spend excessive time on data cleaning and preprocessing. The clarity of the data made it easier to extract relevant information and derive meaningful insights. Additionally, I found Power BI to be a highly useful tool, which greatly facilitated my ability to navigate through the software and create various visualizations. The diverse range of visualizations I generated provided different perspectives and deepened my understanding of the dataset. These visualizations were instrumental in communicating the findings effectively.

Furthermore, the predictive modeling phase went smoothly once I discovered the importance of data transformation using a scaler for non-classifier models. This step helped me standardize the data and ensure optimal performance of the models. With this adjustment, I was able to train my models effectively and achieved decent results. For instance, my logistic regression model achieved an accuracy of 74%, indicating its ability to make accurate predictions, and my multiple regression model yielded a respectable R-squared value of 0.69, demonstrating its ability to explain the variability in the data. Despite a higher mean squared error for multiple linear regression, these results provided valuable insights into the relationships within the dataset.

However, there were certain challenges and areas that did not go as well as expected. One significant limitation was the unavailability of a live data source. This hindered the usefulness of the data, as it lacked real-time updates and excluded recently released movies that may have had updated information. The absence of real-time data limited the timeliness and relevance of the analysis. Additionally, acquiring a sufficient amount of data proved to be a time-consuming task. The code used to retrieve the data took over an hour to iterate through a sample of 20,000 records. This prolonged data acquisition process significantly impacted the scope of the analysis and prevented me from exploring a larger dataset. Moreover, during the data retrieval process, I encountered errors with some IMDb IDs within the loop. I addressed this by implementing a pass exception, which allowed the code to continue execution, but further investigation would be required to understand the root cause.

Looking back, there are several changes I would make given the opportunity. Primarily, I would have liked to utilize the IMDb database, as it offers a more comprehensive and real-time dataset. Access to additional variables, such as movie ratings, would have enriched the analysis and provided deeper insights into movie success factors. Furthermore, I would have liked to explore how movie revenues were distributed across different regions or countries. Understanding whether certain genres generate more revenue in specific countries would have provided valuable information.

For individuals working on a similar project, I would strongly recommend considering paying for the IMDb API or seeking approval to use it. The IMDb database would provide access to a more complete and up-to-date dataset, allowing for a more comprehensive analysis. Additionally, it is essential to work with a dataset that is more normally distributed. This would help avoid potential biases and enable more accurate statistical analysis.

Throughout this project, I have learned several valuable lessons. Firstly, the significance of having reliable and real-time data cannot be overstated. Real-time updates provide more accurate and up-to-date insights, enabling better decision-making. Secondly, effective time management is crucial in data analysis projects, particularly when dealing with large datasets. Finding efficient ways to handle data retrieval, preprocessing, and modeling is essential to meet project timelines. Lastly, adaptability and problem-solving skills are vital when faced with unexpected challenges. Being able to handle errors and find appropriate workarounds is essential to keep the project on track and ensure its success

**Conclusion:**

In conclusion this project presents an analysis of movie revenue, examining the performance of different models and identifying Logistic Regression as the most effective predictor. I found strong positive correlations between revenue and variables such as vote count and budget, highlighting the significance of audience engagement and financial investment in determining a movie's financial success. The ease of use of TMDb API and access to understandable data, along with the ease of navigating Power BI and generating insightful visualizations, were notable strengths of the project. However, limitations included the absence of live data, which hindered real-time accuracy, and challenges in acquiring a sufficient amount of data within a reasonable timeframe.

To enhance future analyses, it is recommended to gather a larger sample of movies, encompassing various genres, release dates, and production budgets. This would provide a more comprehensive understanding of the factors influencing movie revenue. Additionally, incorporating variables such as marketing cost, social media presence, critical reviews, and competition within the same release window could offer valuable insights into other significant aspects impacting revenue. Seeking live data sources would also be beneficial for creating more up-to-date models that reflect the dynamic nature of the film industry and current market trends. These improvements would contribute to more accurate predictions and a deeper understanding of the complex dynamics of movie revenue. In the future, I would like to get access to IMDb API and do an analysis to try to predict an IMDb rating that a movie would have based on similar variables present within this project.

Lastly, despite limitations, the project provides valuable lessons learned, such as the importance of reliable and real-time data, efficient data handling, and adaptability in overcoming unexpected challenges. By addressing the identified limitations and incorporating the recommendations, future research in this area can build upon these findings and further advance my understanding of movie revenue prediction and analysis. In the future, I would like to get access to IMDb API and do an analysis to try to predict an IMDb rating that a movie would have based on similar variables present within this project. These lessons have equipped me with valuable experience and knowledge for future data analysis projects.