Movie Rating and Box Office Return Analysis

Richard Zhu (rz2123) and Brian Yang (bay2006)
New York University

Abstract

With immense catalogs of movies from streaming services like Netflix and Hulu, consumers have unprecedented access to films. In this era of abundant cinematic choices, understanding the factors contributing to a movie's success is paramount. This study aims to gain insight into the dynamics of movie performance, examining the intricate relationships between critical acclaim, commercial success, country of production, and genre. We developed a pipeline to perform ETL on a dataset containing information on over 80,000 movies within the IMDb database and drew inferences from statistical and qualitative analyses. Our investigation revealed that box office return was not a strong predictor of movie ratings. Country of production also proved to be an insignificant feature, as average rating and box office return did not follow a clear trend across countries. However, an examination of genre-specific trends seemed to show a negative linear relationship between average rating and box office return within genres. We concluded box office performance alone was not a significant predictor of a movie's ratings, but a model that included other factors like genre or seasonality could provide more accurate results.

1. Introduction

As internet accessibility continues to grow, positive online ratings are perhaps the most influential factor in drawing larger audiences in the film industry. Most individuals refer to popular review platforms, like IMDb or Rotten Tomatoes, to gain an unbiased estimate of the quality of a film. As such, it is essential for movie production companies to understand the factors that contribute to high ratings. Through our analysis, we expected to find box office returns to be a strong predictor of IMDb ratings. In addition, we explored several other factors we believed to play an important role in movie success. We delved into the relationship between ratings and box office returns with respect to genres and countries of production. Our analysis of genres and countries hoped to provide insight into consumer's tastes. For example, we wanted to understand if genres with a wider appeal, like family or drama, tended to have higher acclaim than genres like horror.

In this paper, we propose a method for analyzing comprehensive movie review data. To navigate the abundance of information online, we leverage big data processing tools such as Spark, Hive, and MapReduce to efficiently store, process, and analyze our data. Our approach provides a framework for an easily scalable pipeline with few limitations and bottlenecks.

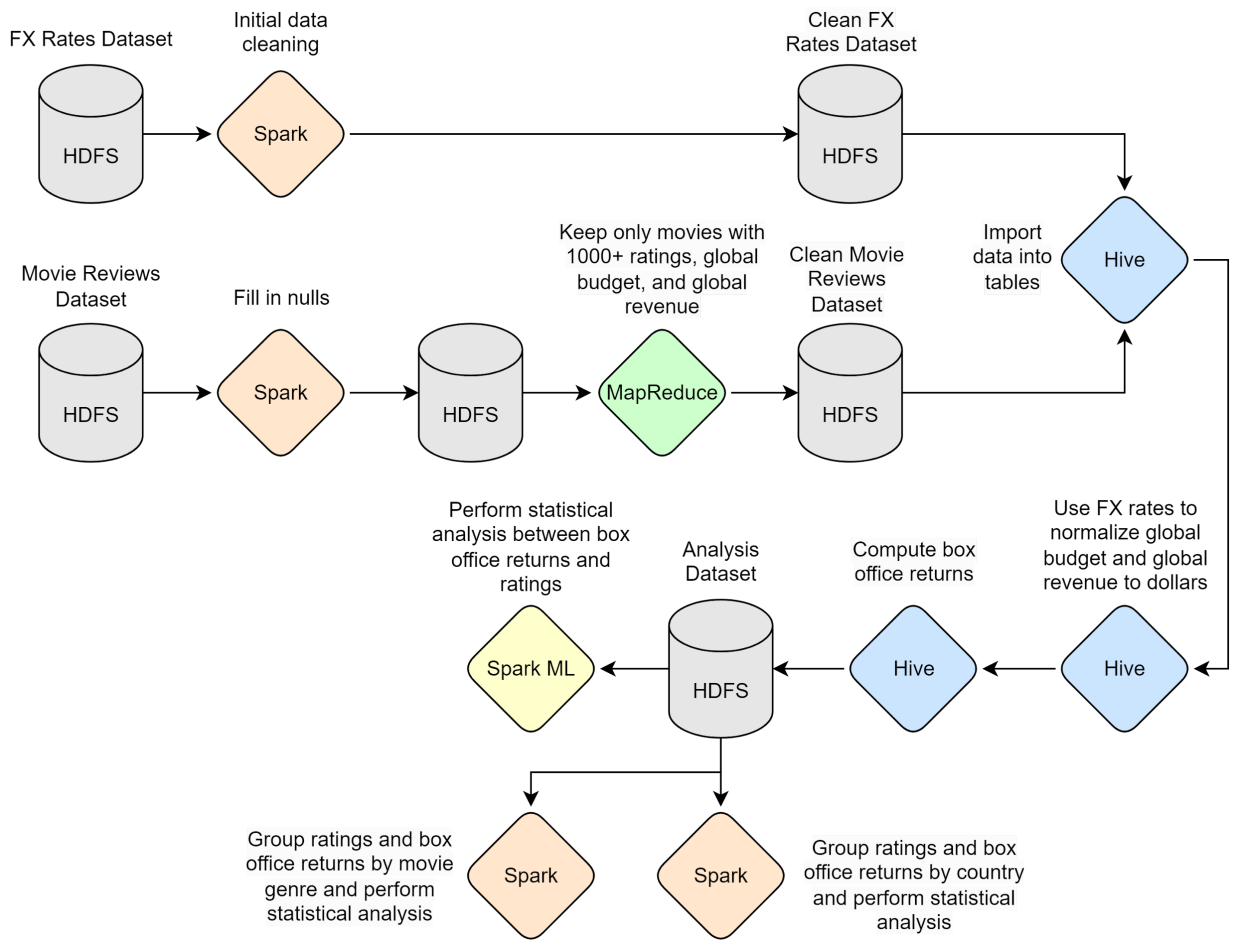

Figure 1: Flowchart representing data processing and analysis pipeline

Our pipeline, pictured in Figure 1, begins with two datasets: FX Rates and Movie Reviews. The FX Rates Dataset undergoes initial data cleaning with Spark, followed by an aggregation step that averages exchange rates for each currency per month/year. Post-aggregation, the dataset is imported into a Hive table. Simultaneously, the Movie Reviews Dataset is processed to fill in null values using Spark, followed by cleaning through MapReduce. The dataset is then refined to include only movies with over 1000+ ratings, global budget, and global revenue figures. This cleaned dataset is then also imported into Hive tables. Next, Hive is used to normalize the global budget and revenue data, using the FX rates to convert foreign currencies into dollars. Spark ML is then employed to perform statistical analysis between the box office returns and ratings. This analysis is further explored by grouping data by movie genre and country for more detailed statistical insights.

2. Motivation

Our research is important because the evolving landscape of the film industry is shifting into an era where data-driven decisions are more impactful than ever. A film's success has historically

been measured by its box office returns, but with the emergence of online rating platforms like IMDb, a different aspect of audience engagement and critique can be observed. Our team is united by a shared interest in the intersection of data analytics and the cinematic world. We conducted this study to discover patterns and empirical insights into the factors of a commercially successful movie.

Each team member brings a different perspective to this study. One of us has a background in data science and is drawn to the application of statistical models to understand the relationship between ratings and returns. Another has a background in financial mathematics and is intrigued by the narrative of numbers behind cinematic success. Our research is driven by our collective curiosity to map out the blueprint of a successful film. The insights from our analysis are intended to inform filmmakers, investors, and actors with data-backed strategies for future endeavors.

3. Related Works

3.1 Movie Box Office Performance Prediction Using Social Media

Apala et al. [3] uses historical data and social media metrics to predict box office performance. Their data – historical movie data and social media metrics – is very similar in nature to our data – historical movie and audience ratings data and historical foreign exchange data. Our analysis aims to discern the extent to which box office performance can predict IMDb ratings and serve as a valuable tool for movie production companies in understanding consumer preferences and trends. Just like their study, we also look at extraneous factors. In our case, we further assess genre and country; in their case, they look at marketing and star power. Since our data and the analytics derived from them are very similar to the data and processes described in the research paper, the credibility and validity of our study are strengthened.

3.2 Movie Performance Prediction From Plot Summary and Character Description

Lee et al. [2] attempt to predict the success of a movie, as determined by its score on Rotten Tomatoes, by developing deep learning models that analyze the movie's summary. They take a different approach to their analysis, as the main focus of their paper is the development of a natural language processing model that processes movie summaries. Instead of trying to predict a movie's success, we are more focused on what factors have the highest predictive power for success. This paper focuses on the summaries of the movies, while we explore factors like genre and country. Their model yielded accurate results for certain movie genres, so it seems that we may gain additional insight with our analysis.

4. Datasets

4.1 IMDb Movies Dataset

The IMDb Movies Dataset [4] was uploaded to data.world on 9/14/2020 and contains data on 80,000+ movies from the online IMDb database. The movies' release dates range from 1906 to 2019. The relevant fields for each movie in our analysis are title, date published, genre, country, average IMDb rating, number of ratings, budget, and worldwide gross income. The budgets are listed in various currencies – typically the local currency in which the movie was produced. Budgets in USD are listed with a dollar sign followed by the amount, e.g. "$ 1000". Budgets in other currencies are given with the currency ticker followed by the amount, e.g. "EUR 1000". Worldwide gross income is always given in USD and follows the same convention as the budgets.

4.2 FX Rates Dataset

The FX Rates Dataset [1], last updated on 05/27/2021, compiles historic foreign exchange rates between 01/1999 and 05/2021 from the European Central Bank. It includes rates for various currencies with USD as the quoted currency (x/USD). The FX The dataset is 3.2 MB and includes the following columns: Date, EUR, JPY, BGN, CZK, DKK, GBP, HUF, PLN, RON, SEK, CHF, ISK, NOK, HRK, RUB, TRL, TRY, AUD, BRL, CAD, CNY, HKD, IDR, ILS, INR, KRW, MXN, MYR, NZD, PHP, SGD, THB, and ZAR.

5. Analytics Stages

5.1 IMDb Dataset Preparation

5.1.1 ETL

We began the ETL process on the IMDb movies dataset by uploading the raw CSV file to HDFS from a local computer. This dataset was fairly sparse, especially for the budget and global revenue columns. We created a Scala script that filled all blank records with the string literal "Null", for better compatibility with MapReduce, and saved the result back in HDFS. We then cleaned this dataset with MapReduce, applying several transformations to filter our data. We only kept movies in the dataset with more than 1,000 IMDb ratings and with budget and global revenue. The budgets were specified in local currencies, so the budget column was split into two: one for the currency ticker (e.g. USD, EUR) and one for the numerical value. Finally, we dropped any unnecessary columns for our analysis and saved the resulting data as a CSV file within HDFS.

5.1.2 Profiling

Before cleaning the dataset, we took note of the number of unique values within the columns for IMDb rating, title, and year. We performed this with a MapReduce job, where the Mapper emitted a key-value pair of the column name as the key and the data within that column as the

value. The Reducer received these pairs and determined the unique count of values within each column.

After the data was cleaned in the ETL process, we performed more advanced analysis on some of the numerical columns. We found the mean, median, mode, and standard deviation for columns like budget and global revenue. After cleaning, we once again determined the unique values for rating, title, and year with the same MapReduce job as before. We found the values for rating and year to be relatively similar to the initial values, but the unique values for title drastically dropped. This matched our expectations, as there still should have been a wide range of ratings and years, but all the movies that didn't have sufficient data were dropped.

5.2 FX Rates Dataset Preparation

5.2.1 ETL

To begin the FX Rates Dataset ETL process, we uploaded the raw FX Rates CSV files to HDFS from a local desktop. The raw CSV file was then loaded into an RDD in Spark-Scala. We removed the header and first row, which contained unnecessary column names and data. After this initial clean-up, the data was gathered and written to HDFS in preparation for the next phase of transformation.

For data transformation, the cleaned dataset was re-loaded into Spark-Scala. We applied a custom function to each row, splitting the date into year and month components that were appended in new columns. Following this, the header was adjusted to reflect the new schema, now including separate columns for year and month. The transformed data was then consolidated into a single partition and saved back to HDFS, resulting in a clean and structured dataset ready for in-depth analysis.

5.2.2 Profiling

The profiling of the FX Rates dataset began with the loading of the raw CSV data into Spark-Scala. We then computed the total count of the records to understand the dataset's size. Using a map operation, we transformed each row into a key-value pair, where the key was the date, and the value was the concatenated string of currency values. The distinct values in each column were counted to assess the diversity of the data.

For the statistical analysis, we proceeded with the cleaned CSV data produced from the ETL process. Each line was split into columns, discarding the date column as it was no longer necessary for the analysis. We generated separate RDDs for each currency column, which were then used to compute statistical measures. Our custom computeStatistics function was applied to calculate the mean, median, mode, and standard deviation for each currency column. The function handled NaN values to maintain the accuracy of our statistical outputs. The computed statistics for each column provided insights into the central tendency and dispersion of the exchange rates.

5.3 Ingestion

For data ingestion, we imported the cleaned datasets into Hive by creating external tables, allowing the data to be stored in HDFS but managed by Hive. The clean movies and fx rates datasets were imported into tables using OpenCSVSerde for CSV format compatibility. To improve compatibility between the two datasets, we computed average monthly exchange rates. We then used a join operation between the movies and fx tables, keyed on the year and month fields. During this join, an `exchange_rate` field was created by retrieving the exchange rate for each movie's currency. This produced a joined table comprising all fields from the movies table, along with the associated average exchange rate given the movie's date of release. Subsequently, we multiplied the original budget and revenue figures with the exchange rate to create a normalized joined table, which stored all figures in USD. A new field, `boxoffice_return`, was also computed to indicate the return on investment, calculated by dividing the global revenue by the normalized budget. This field, critical to our analysis, represents movie box office returns as a percentage. The final step in our ingestion process involved exporting this normalized movie data to a directory on HDFS in CSV format for further analysis.

5.4 Analytics

5.4.1 Rating and Return Analysis

In the analytics phase of our study, we employed Spark's machine learning library to predict IMDb ratings based on the box office returns of movies. We first prepared our dataset by using Spark's VectorAssembler to transform the `boxoffice_return` feature into a vector, the required input for Spark's ML models. We then split our data into an 80-20 train-test split for model validation purposes. Two regression models, Linear Regression and Random Forest Regression, were trained on the training set. Upon evaluation using the Mean Absolute Error (MAE) metric, the Linear Regression model yielded an MAE of 0.806, while the Random Forest model achieved a slightly better MAE of 0.777.

The results indicate that both models have a similar ability to predict movie ratings from box office returns, with the Random Forest model showing a marginally superior performance. Considering the MAE values in relation to the rating scale, the errors suggest that the models have a moderate prediction capability, given the rating scale is 1 to 10. The MAE implies the predictions were, on average, around 0.8 rating points away from the actual values. This relatively high MAE suggests that factors other than box office returns may have a significant influence on movie ratings.

5.4.2 Genre and Country Analysis

The second portion of our analytics focused on finding the relationship between a movie's genre and country of production with its success. Our metrics for success were a high IMDb rating and a high box office return. To perform this analysis we utilized Spark, loading our normalized data from HDFS into a DataFrame. Movies often had multiple genres and countries listed, so we

exploded on the genres and countries columns. After this, we grouped the movies by either genre or country and determined the average IMDb rating and box office return for each group. We only wanted to keep groups with a significant amount of data, so we dropped any groups that had less than 30 data points.

After finding these averages when grouped by country and genre, we exported this data and created visualizations with Matplotlib in Python. Our investigation into the country of production, shown in Figure 2, showed no clear structure. Our genre-specific analysis, shown in Figure 3, revealed an interesting pattern. Overall, there seemed to be a negative linear relationship between ratings and box office returns across genres. On one hand, fictional genres, like horror and family, exhibited higher returns and lower box office success. On the other hand, non-fictional genres, like history and war, demonstrated lower returns and higher box office success.

6. Visualizations

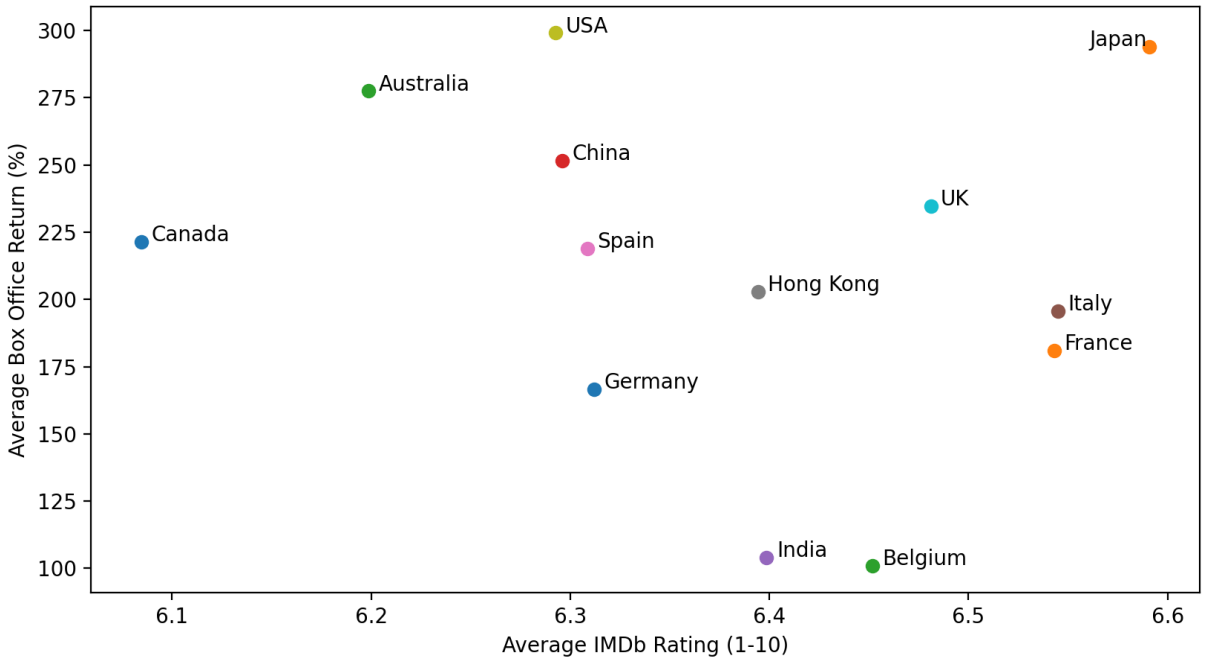

Figure 2: Average box office returns and IMDb rating for movies grouped by country

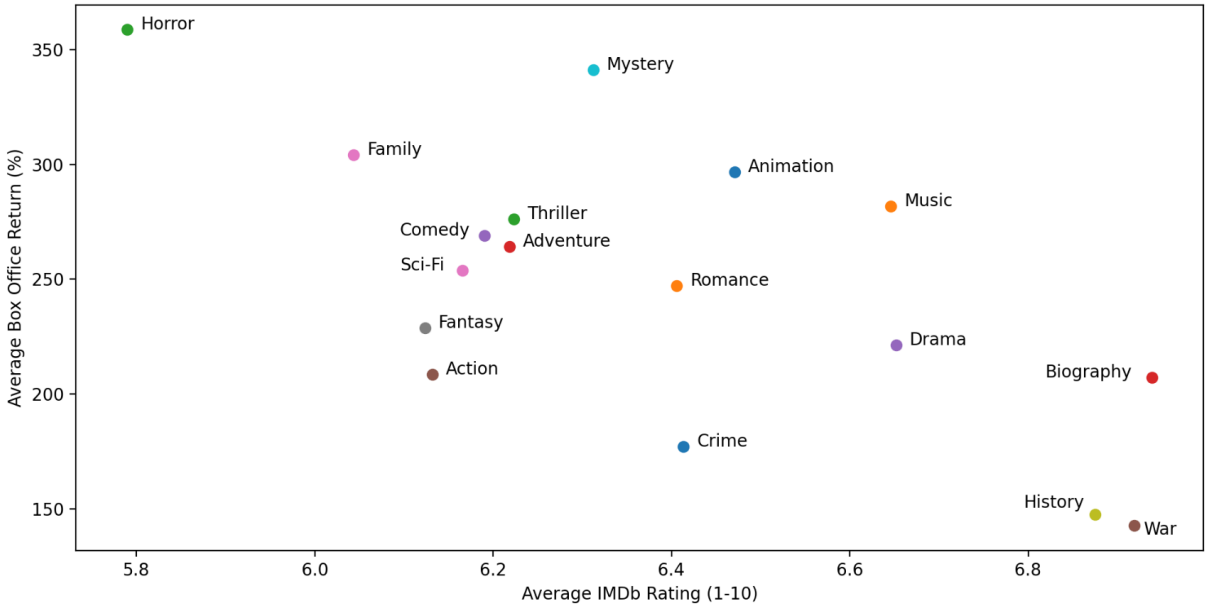

Figure 3: Average box office returns and IMDb rating for movies grouped by genre

7. Conclusion

Contrary to our initial hypothesis, our analysis demonstrated that box office return was not a powerful predictor of IMDb ratings. This finding shows the need for a more nuanced understanding of the factors that contribute to a film's reception beyond conventional measures of financial success. Our analysis of genre, however, yielded interesting results. We hypothesize that our results can be explained by generalized assumptions of genres and viewership. Non-fictional genres, like history and war, likely attract viewers who have prior interest in the subjects portrayed. Due to the niche nature of these genres, the result is a relatively small but engaged audience. On the contrary, fictional genres, like family and mystery, typically feature universally entertaining stories. These narratives appeal to a wider audience, but often lack thematic depth. This potential difference in viewership could explain the trend we observed in our visualization.

It is evident from our findings that creating a well-rated movie involves navigating factors beyond box office return. Industry professionals must consider a holistic approach, considering the impact of extraneous variables that contribute to a movie's overall success. Our study lays the technical framework for future research, allowing others to extend the analysis. By incorporating additional features such as social trends, seasonality, and age rating, a more comprehensive understanding of movie rating performance could be achieved.

8. REFERENCES

[1] Humanitarian Data Exchange. (2019). Foreign Exchange Rates [Dataset]. humdata.org. https://data.world/hdx/3c7aca84-e00b-46a0-a8d5-d769b16e5351

[2] J. -H. Lee, Y. -J. Kim and Y. -G. Cheong, "Predicting Quality and Popularity of a Movie From Plot Summary and Character Description Using Contextualized Word Embeddings," 2020 IEEE Conference on Games (CoG), Osaka, Japan, 2020, pp. 214-220, doi: 10.1109/CoG47356.2020.9231541.

[3] K. R. Apala, M. Jose, S. Motnam, C. . -C. Chan, K. J. Liszka and F. de Gregorio, "Prediction of movies box office performance using social media," 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013), Niagara Falls, ON, Canada, 2013, pp. 1209-1214, doi: 10.1145/2492517.2500232.

[4] Mahesh. (2020). Movies [Dataset]. data.world. https://data.world/mahe432/movies