Myles J Sartor

Wei Ai

INST447-0101

16 December 2025

Final Project Report (IMDB)

Introduction

Films play a significant role in shaping culture, entertainment, and public discourse throughout the world. Over the past several decades, the rise of digital platforms has continuously transformed how audiences choose to consume, evaluate, and even discuss movies. Platforms such as IMDb allow millions of users to rate and review films, producing large-scale datasets that reflect collective audience opinion. These datasets provide valuable insights into what makes a movie popular or commercially successful in the eyes of the public. Understanding what ultimately drives a film's success is important. In providing perspective into genre performances, runtime preferences, and audience engagements, this information can inform production and marketing decisions so that future products can be worthwhile. Furthermore, film data allows one to realize how cultural products are evaluated by broad audiences. The use of public APIs and structured datasets makes film analysis a practical and meaningful domain to branch into. This is especially true for stakeholders such as film critics, movie directors, and Hollywood companies seeking to maximize profits on what they choose to release and invest in. This reality raises questions about factors that are associated with higher IMDb ratings and greater audience engagement for films. My project delves into these details for cinema released between the years 2000 and 2022. It was imperative to uncover if certain film genres consistently received higher IMDb ratings than others (which changed over time on average), and if movie runtime was associated with higher or lower audience ratings. It was also extremely beneficial to understand how audience popularity (measured by IMDb vote count) related to box office revenue. After combining a large IMDb dataset with supplemental data from an open movie database API (OMDb), Python pandas and matplotlib were utilized to perform data cleaning and exploratory data analysis for valuable interpretations to be made. In working with the JSON format, the IMDb dataset had movie titles, IDs, release years, genres, and more regarding films from the years 2020 to 2022. The API was a REST API (JSON format) that held additional metadata that could uniquely identify a film when paired with other information (Quasi). The use of both allowed multiple data sources to be merged into a new data frame format that fit the scope of the finalized product.

Dataset Description

The IMDb dataset contained several thousand films released between 2022 and 2022, providing a substantial volume of data for exploratory analysis. Since so much was available, it was much easier to conduct meaningful aggregation by genre and year alongside rating distributions and popularity metrics. However, due to API rate limits imposed by OMDb, only data from the first 200 movies present was utilized. This still proved to be sufficient for the purpose of identifying general patterns and relationships that existed originally in this domain. Being a part of a 22-year period meant that these trends over time were much more intuitive to find, despite the data's static nature (Velocity). The clear evolution of audience ratings given the different changes in the industry was clear to see, as they lined up with the events of history. For instance, when streaming platforms started to get more popular among the general population, a fluctuating change in the way movies were rated was quite visible. The same could be said for films released around the time of the pandemic in 2020. When it came to engaging with different types of data, the sheer amount of diversity present promoted levels of structure in all forms of analysis. Numeric (Int) and categorical (String) variables worked together to exist in complex formats that were stored in dictionaries (ratings) and lists (genres). This level of variety required significant data preparation that used techniques such as JSON flattening, list exploding, and conversion of text-based numeric fields. Regardless, these attributes did not prevent data quality and reliability issues from exposing themselves later on during data manipulation. There were often missing or incomplete runtime and box office values that disrupted the accuracy of the visualizations if not handled prior to plotting. At other times, there were inconsistent genre labels, API responses, and potential bias in IMDb ratings that differ from the general audience's opinion. If not for filtering, type changes, and careful operations, they would've remained as limitations affecting the certainty of resulting conclusions. This movie data is valuable in its true essence, as it's rare that direct comparisons between engagement, commercial success, and general perceptions can be drawn so reliably (Valuable and veracious). With all its expansive information, the research question at hand can be definitively explored and answered in a meaningful way.

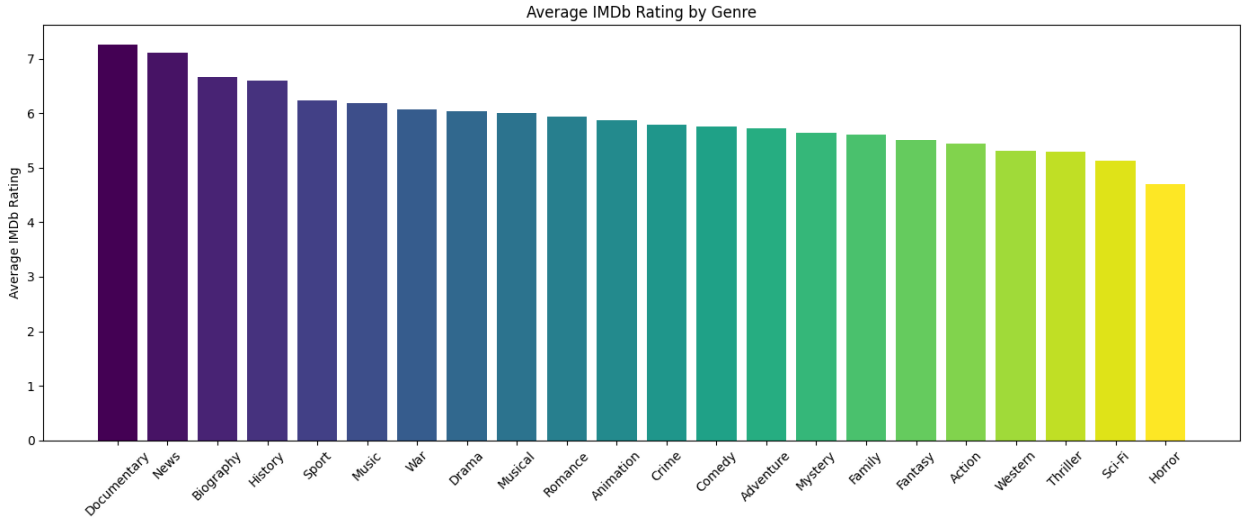
Data Preparation and Exploration Process

The IMDb data was stored in a JSON file that was parsed line by line after importing the JSON package. Every line was placed into a Python dictionary and stored in a list, which was then transformed into a pandas data frame. The OMDb API was accessed using a function that had a GET request for each IMDb ID while being capped at 200. In addition, a delay was placed so that rate limits were not reached eventually. If the response wasn't successful, then it didn't go through, preventing invalid responses from being present in the data. The two datasets were merged using IMDb IDs with a left join, ensuring that all IMDb movies remained in the final dataset even when OMDb data was unavailable (only matches returned from the OMDb data, while everything from IMDb data was present). In cleaning the data, it was very important for the ratings dictionary to be flattened into separate numeric columns to enable further aggregation

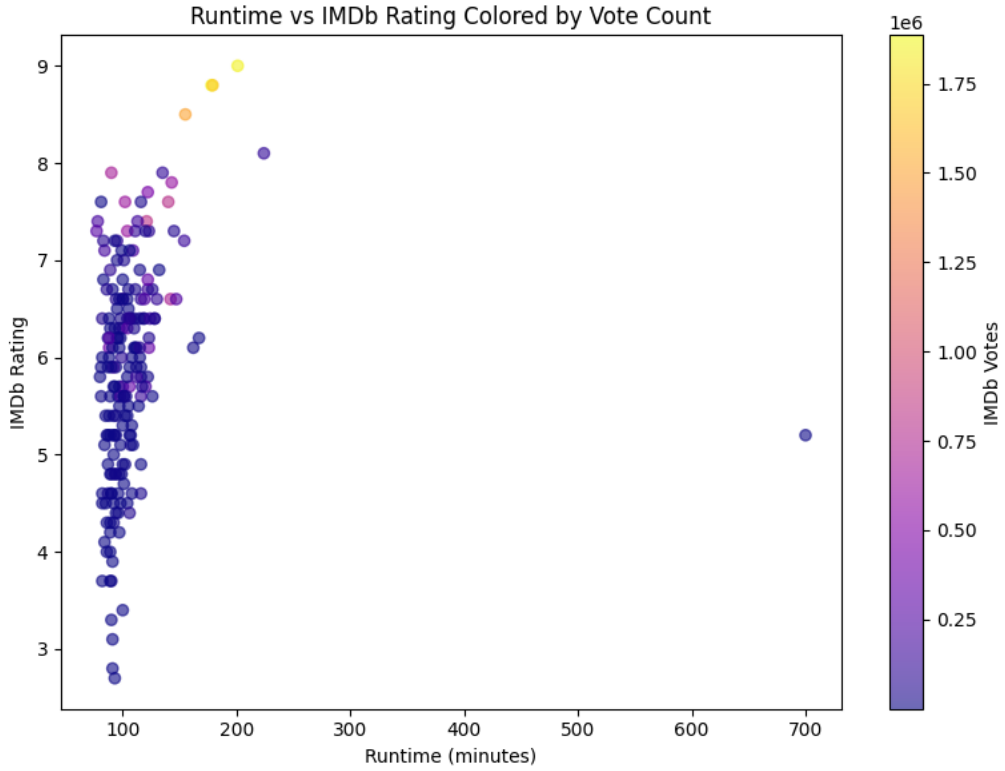
and visualization. Without this, average ratings and vote count could not be deciphered properly prior to analysis. Runtime values included text labels with phrases like “120 min”, which required removal prior to conversion to numerical values for statistical testing. Box office values held dollar sign symbols and commas that needed to be stripped and replaced before conversion to numeric form as well. When it came to genres, the lists were efficiently exploded so that each genre could be analyzed independently once empty or missing genres were removed if invalid. After all this, it was possible to go from raw semi-structured data to formatted data suitable for examination. The exploratory data analysis focused on understanding the distributions given. From missing values to key variables, grouping and aggregation were utilized extensively to compare genres, investigate average ratings, and analyze popularity metrics. Visual explorations played a crucial role in identifying patterns and guiding interpretations.

Analysis and Findings

The first visualization is a bar plot that looks at comparing IMDb ratings across genres with a specific subset of data that was merged and pulled together from data processing efforts. The x-axis holds the genres, and the y-axis encompasses the average ratings, letting one determine if the average rating effectively depends on the genre that the film came from. The code specified a look at only the top 10 genres for relative success, and a gradient-based coloring system was utilized for the length of the space that the genre data frame took up (Index). Genres such as documentary, news, biography, history, and sport exhibited higher ratings, while action, western, thriller, sci-fi, and horror scored lower on average. This pattern suggests that audiences evaluate genres differently, often applying stricter standards to entertainment-focused genres and more favorable evaluations to narrative or informational films. However, this is likely because it can be harder for film directors to reliably execute movies from these genres. Usually, the material in them is harder to work with, and people aren’t as willing to sit through lackluster work from these mediums if they deal with such topics. An honest effort in more educational categories goes a long way in the eyes of many average consumers.

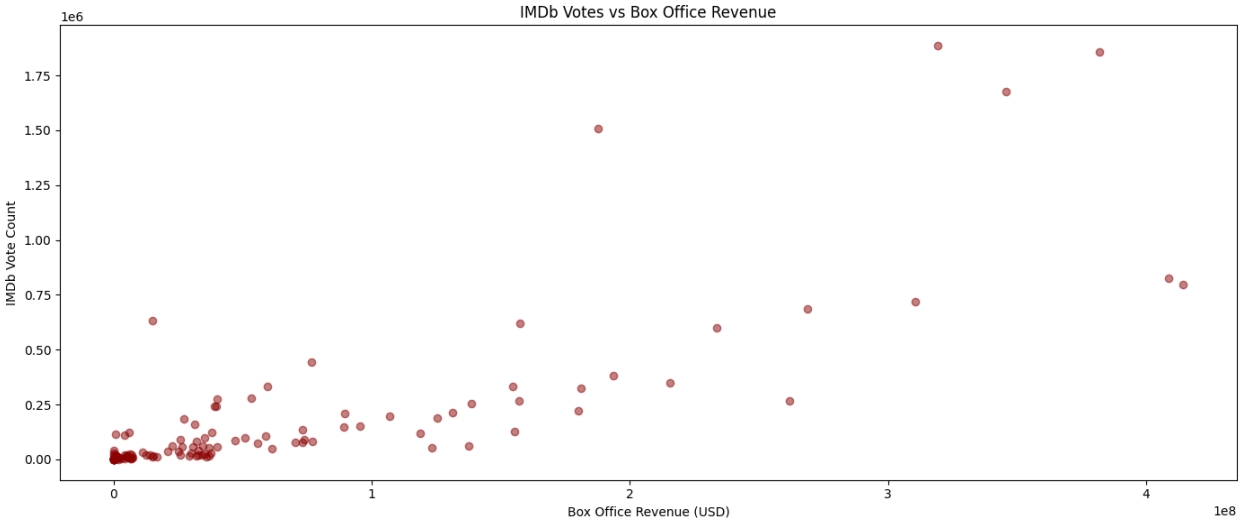


The second visualization was a scatterplot that examines the relationship between runtime and IMDb rating, with gradient-based coloring that indicates changes in vote count. The color bar and governing results show no strong positive or negative relationship between runtime and rating. Ratings are out of 10 on the y-axis, and runtime is scaled using minutes on the x-axis to see if the rating depends on the runtime. A few relatively shorter films showed low ratings of around 3/10, but most of the output congregated around the middle, from ratings 4-7. One movie that was 11 hours received only a rating of 5/10, while a few movies that were over 2 hours long received exceptionally high ratings, around a 9/10. The IMDb vote outcome reveals that popular films span a wide range of runtimes, reinforcing the idea that runtime alone does not drive audience engagement or approval. This lines up with what one might expect, since people often evaluate a movie on its content and quality rather than just how long it is.

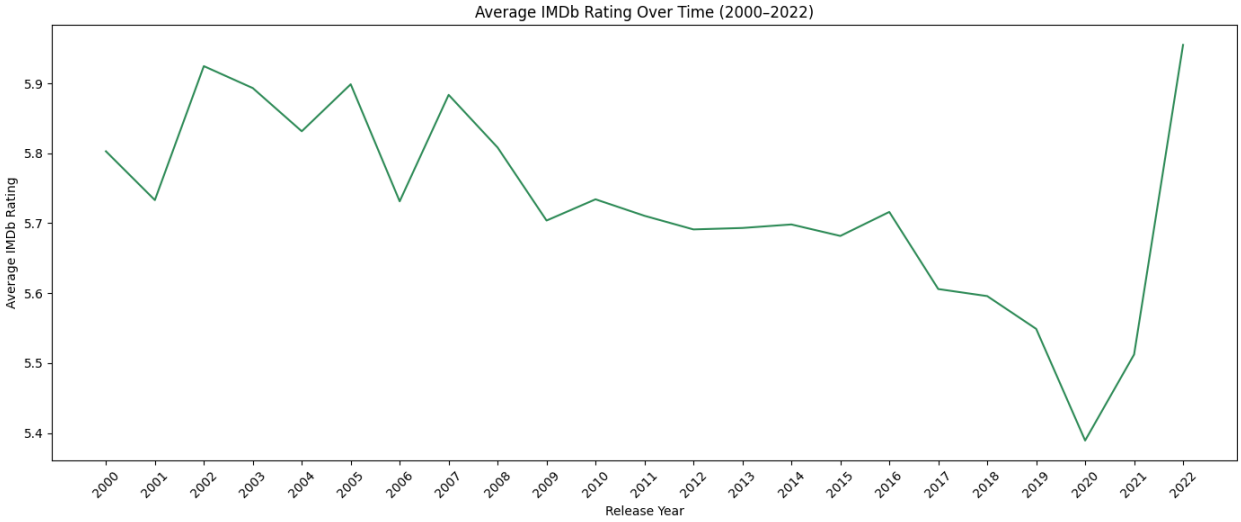


The third visualization reveals the relationship between box office revenue and IMDb vote counts through the use of a simple scatterplot. The revenue on the x-axis is on a base scale, and labeling of up to 100 million (meaning the label 1 is 100 million), and the vote count within the y-axis is on a base scale, and labeling of up to 1 million (meaning the label 1.00 is 1 million). The primary goal was to determine if the vote count depended on the financial successes of whatever was released. The large bulk of the movies congregated around the lower end of the graph, with revenue returning up to around 40 million to 100 million USD. In showcasing a positive correlation between the two, films with higher earnings tend to receive more votes. Although the relationship isn't particularly strong (Likely due to sample size), it was still significant since increased visibility and broader distribution channels for popular cinema usually

increase sales. This finding highlights the distinction between popularity and quality, as commercial success increases public engagement but does not necessarily imply higher ratings; it's simply easier to find. For instance, there are a select few instances in the plot where box office revenue appears to be lower, but the vote count is higher than in other places where the revenue was higher, and the vote count was lower.



The final visualization tracks average IMDb ratings by release year with a line plot that temporally analyzes the relationship from 2000 to 2022 with modest fluctuations. The release year on the x-axis and the average rating on the y-axis serve to determine if release years affect ratings in any capacity. From 2000-2007, this varied, going up and down (oscillating) based on the current climate of the film industry. Then from 2007-1016 there was a sharp and then steady decline that plateaued before shortly rising. During the time period of 2016-2020, there was another sharp decline. This was likely brought on by the lack of sufficient quality movies leading all the way up to the pandemic, where movies weren't being produced as often. However, despite this downward trend, things have since recovered as ratings recently went back up to their normal levels from 2 decades prior. The graph tells people that the perspectives of those who view cinema are constantly changing. Whether they decide to participate in the critique of film more or not, ratings get distributed on a broader scale over time, and the pathway for success is never truly straightforward.



Conclusion

Film success is influenced by multiple factors as long as audience members continue to remain in control. Genre plays a meaningful role in shaping audience ratings, while popularity and commercial success are closely related (similar to other art mediums). However, metrics like runtime and release year tend to display less straightforward and weaker relationships when it comes to any given movie. The limits placed on the sheer size of the data being analyzed due to API sampling limitations and missing data certainly lower some of my confidence in these findings. However, the selection of data that I obtained prior to further processing was sufficient for some valid conclusions to be drawn from said analysis. The patterns observed were consistent with reality and general expectations about audience behavior and media consumption. From an industry perspective, the results suggest that audience evaluation being genre dependent means genres associated with real-world narratives or informational content, such as documentaries or biographies, benefit from lower audience expectations around entertainment value. Instead, they'd rather have directors delve into storytelling, authenticity, and educational impact, which are easier to address and keep people engaged with for a good rating. Filmmakers and studios operating in these genres usually prioritize critical acclaim and long-term reputation over immediate box office performance. In other words, they're more focused on getting things right rather than trying to capitalize on the monetary gains that come with making a movie. On the other hand, genres such as action and comedy appear to be judged harshly by audiences due to the high expectations for amusement and a sense of enjoyment from elements like pacing, cast, and spectacle. If they do not deliver on these promises, the overarching outcome will not be favorable. The relationship between box office revenue and IMDb vote counts highlights the role of visibility and distribution in shaping audience engagement. The number of people in the market that one can reach significantly influences how many users rate a film positively or negatively. This reinforces the importance of strategies that go towards generating audience participation and online engagement metrics. The lack of a clear relationship between runtime and rating indicates that audiences do not systematically prefer shorter or longer films. Instead,

the content quality and genre expectations likely play a more important role in determining what sticks with somebody after viewing the movie. This insight challenges the common assumption that longer films are inherently better received because of the work that is seriously put into them. Instead, people would rather the narrative effectiveness take precedence over a work's duration. Although all of these metrics provide verifiably different dimensions of success that are reliable enough for valuable conclusions to be drawn on their own, future work would likely look into an expansion of API coverage past only 200 films, since a larger sample would increase my overall confidence in the output. It would also make sense to look into predictive modeling techniques that use machine learning to predict IMDb ratings or vote counts based on features such as genre, runtime, release year, box office revenue, and critic scores. In utilizing supervised learning, the analysis would become substantial, as it would have use cases beyond being descriptive of the environment. Including streaming platform metrics could also aid in capturing shifts in film consumption over time, as people switch from indulging in theatrical releases to engaging with streaming services (noting differences across platforms, especially during the COVID-19 pandemic).

Reflections on Tools and Process

AI tools were utilized strategically throughout different stages of this project to help with brainstorming topics and integrating some of the data from them. This helped narrow the scope of the project down to meaningful standards that aligned with the analysis I planned to do in the future. During the implementation of the technical methods, AI was particularly helpful for debugging code or explaining a procedure that I didn't quite understand when it came to manipulating JSON data, similar to how we had previously done it in class. For example, flattening nested rating dictionaries and exploding genre lists for analysis was a topic that we had only briefly covered in assignments throughout the semester without going into much detail. In these situations where I felt the methods would be useful, I was able to use AI to bolster my understanding and refine my data for visualization choices and plot interpretation clarity. Problem-solving became much simpler during data processing, and decoding errors while merging and mapping colors to certain portions of my work were more intuitive. However, AI was less useful when it came to interpreting potential results and actual visuals to use in context. It wasn't able to fully replace any of my human judgment that could accurately assess what was actually causing findings to show up the way they did in this film domain. Without critical thinking, certain correlations and unsupported claims it threw out after going through its own data limitations were essentially unusable. Through this, the project was able to teach me the importance of making sure your data is viable prior to using it for future analysis. The bulk of the work for many data science projects is done during the cleaning, merging, and transforming of unstructured data from multiple sources before anything meaningful is drawn from it. Data in the real world is rarely up to standards for proper exploration. In having the API to use as a backup tool and visuals to use as an analytical tool, I was able to tell a "story" with the data that led to insights you couldn't simply see with just the numbers in front of you. As a result, this project

has effectively shaped how I approach data science tasks going forward, as the importance of data quality and missing data can often be overlooked. I was forced to be mindful of the dangers that come with using APIs carelessly, alongside the troubles that come from trying to successfully merge them with a rather large JSON file. The data science workflow is often unique and full of technical challenges that wrestle with domain knowledge and reliable insights.