What Makes a Successful Movie: an Analysis of Box Office Stats and Reviews of Marvel Movies

[*]A paper by the group ”Rizz Kings” for the ”Data in the Wild: Wrangling and Visualising Data (Autumn 2023)” (KSDWWVD1KU) course at ITU.

Adam Lass
adala@itu.dk

Aidan Stocks
aist@itu.dk

Christian Hugo Rasmussen
chur@itu.dk

Oleg Jarma Montoya
oleja@itu.dk

Abstract—**In this report, we are exploring the driving factors of box office success. Using sentence embeddings, hierarchical clustering and linear regression, we try to uncover statistically significant patterns illuminating the relationship between a movie's contextual characteristics and commercial success. Here, we find that budget, franchise status and the percentage of admiring critics are crucial in driving a movie's financial success. However, we also acknowledge the difference between financial success (revenue being higher than the total budget) and the complex metric that is success in general. Despite the insights gained from our analysis, the complexities of screenplays and the limitations of publicly available data pose challenges in identifying the absolute recipe for box office success. Nonetheless, our research still sheds light on valuable statistical insights, builds a dataset and proposes a generic method of clustering movie franchises based on cast similarity.**

I. Introduction

Since the dawn of the internet, people have gathered online to discuss movies they love, hate, or tolerate. Nowadays much of that online discourse is directed through a list of select online movie databases. These sites are commonly used to justify why your movie is (factually) better than whatever movie your friend is trying to make you watch. But does a high score on *Rotten-Tomatoes* or *IMDb* yield more box office revenue? Are all these online discussions just simply a loud minority projecting their agendas, or can these sites actually help indicate the success of a movie?

With this project, we seek to explore what defines a successful movie using only descriptive data such as online reviews and metadata. We have chosen to work specifically with movies within the *Marvel-Cinematic-Universe* (*MCU*) since they all have a large number of reviews. They also provide us with a more homogeneous collection of movies that are inherently similar to each other in genre, studio and budget.

Keeping this in mind, we end up with the following research question: **By analysing movie metadata and online reviews, can we uncover the key elements contributing to a movie's success?**

II. Data Collection

Before answering this question, we must define what we mean by movie metadata and online reviews.

When referring to movie metadata, we mean data such as the budget spent to produce the movie, the revenue made, or the cast of actors, and not data such as the movie's script or length. Luckily, for getting movie metadata, an API from *The Movie Database* (*TMDB*)[23d] was available for use. *TMDB* is a community-built movie and TV database. People can sign up to the website and contribute to the database to make such data easily accessible to the public. Using their API, we were able to acquire the movie metadata needed.

Online reviews refer to reviews left by users on websites such as *IMDb*, where critics and regular audiences can leave a score and written review about their opinion of a movie. To have a variety in data sources, three different websites, namely *IMDb*, *Rotten Tomatoes* (RT), and *Metacritic*, were selected to be scraped and later combined into a single dataset. They were selected because they meet our requirements of allowing critics and users to rate and review movies and because of their popularity and high usage[23f].

The next two subsections describe how the *TMDB* API was used to get movie metadata and our scraping methods for getting movie reviews.

A. Movie Meta Data

Column name	**Sample data**
title	The Marvels
imdb_id	tt10676048
release_date	2023-11-08
cast	{Samuel L. Jackson: Nick Fury, ...}
direction	['Nia DaCosta']
production_companies	['Marvel Studios', ...]
budget	274800000
revenue	108998133

TABLE I
Data variables collected using TMDB API with a sample row of data. ”...” denotes that there is more data in the list/dictionary than shown in the sample.

Using a Python wrapper for the *TMDB* API[23e] along with its labelling system, we could filter the movies of interest and select all the variables we deemed helpful or interesting for the project. The *TMDB* labelling system groups movies depending on certain criteria, a particularly useful label for this project is the *MCU* label, which groups any movie, short or TV show made by *Marvel* studios or at least partially owned by it.

Once *MCU* movies were specified, we filtered through what it returned to get the data columns we deemed relevant for us.

Table I shows the data columns gathered for each movie along with a sample of values from the movie *The Marvels*.

B. Movie Poster Data

As a part of our research, we wanted to analyze the characteristics of movie posters. Here, we made the *TMDB* API save the poster's URL ID of each movie so it can be extracted and later, using parsing techniques, downloaded as a PNG file.

C. Movie Review Data

Since we struggled to find free available datasets and APIs on movie reviews, we manually scraped reviews from popular sites. We could have chosen to work with a pre-existing IMDb dataset. Still, this dataset was deemed dated and contained a relatively low amount of MCU reviews for our purpose.

To save time and split the workload, a different person was assigned to scrape each of the three websites, leading to different scraping implementations.

Due to the dynamic web content in the review pages of the websites, Selenium had to be used in all scrapes. However, the methods used to parse the HTML of pages differed. **Table II** shows, for each website, how Selenium dealt with dynamic content and the parsing method used.

	IMDb	**RT**	**Metacritic**
Selenium use	Scroll Down & Click "Load More"	Click "Next"	Scroll down
HTML parsing	Scrapy	Beautiful Soup	Scrapy

TABLE II
THIS TABLE SHOWS, FOR EACH WEB SCRAPE, HOW SELENIUM WAS USED AND THE METHOD OF HTML PARSING.

Note that reviews without text were skipped during the scraping process due to the interest in analysing review scores with respect to their text. For scraping *Metacritic*, Scrapy was initially selected to scrape a user's average review score based on the usernames of the initial scrapes. However, due to the ethical concerns of overstepping the robots.txt guidelines, this avenue was not explored, leading to the full potential of Scrapy being left unutilized.

Because of the differences in scoring systems used, when combining the data from all scrapes into one dataset, the scores needed to be first transformed to have a standardized scoring system. Each site's different ranges of scores are shown in **Table III**.

RT Critics is a special case as critic scores are hidden and only represent whether a critic considered a movie as "Fresh" (Good) or "Rotten" (Bad).

Ignoring *RT Critics*, the scoring scale used by *Metacritic* does not match the other sites as *Metacritic* uses an 11-point scale instead of a 10-point. To match the scales, the *Metacritic* scores are altered so that the 0 scores become 1. Since the ratings are a categorical system, and 1 is the lowest score in both *IMDb* and *RT*, we only lose some information from the *Metacritic* "1" scores. While this conversion is not optimal it retains some of the information.

Source	**Possible Scores**										
Metacritic	0	1	2	3	4	5	6	7	8	9	10
IMDb		1	2	3	4	5	6	7	8	9	10
RT Audience		0.5	1	1.5	2	2.5	3	3.5	4	4.5	5
RT Critics	Rotten					Fresh					

TABLE III
THE POSSIBLE REVIEW SCORES GIVEN BY USERS ON DIFFERENT WEBSITES. NOTE THAT RT CRITICS HAS A BINARY SCALE AND THAT THE LEFTMOST IMDB AND RT AUDIENCE COLUMN IS EMPTY TO SHOW A MISMATCH IN POINT SCALES WITH METACRITIC.

After the scales were matched, the *RT* Audience scores were multiplied by 2 so that all scales and representing values matched between sites.

To keep RT Critic data for later analysis, the binary Rotten/Fresh scores were stored in a separate column from other Review Scores. Null values were used to differentiate between review scores and rotten/fresh scores should be used.

After these transformations, the scraped data from the three sources were combined into a single dataset for our analysis.

1) GitHub file structure: All data used in this project can be found in our GitHub repository[23c] under the data folder with the following file structure:

```
data/
 ├─processed/
 │  ├─all_reviews.csv ...... All combined reviews
 │  └─movie_stats/
 │     └─movie_stats.csv ........ Movie Metadata
 └─raw/
    └─images/
       ├─tt0371746_iron-man.jpg
       └─ .....................jpg files for all posters
```

III. METHODS

A. Review Embeddings

When analyzing the scraped reviews, we wanted to delve deeper into the written reviews we had collected. The goal being to measure correlation between their scores and the written text.
The method used to quantify the written language within each review is based around sentence embeddings[RG19], which we hoped would help us separate reviews with a "bad" score from ones with a "good" score (in this context, we define this as a score more or less than $\frac{1}{2}$ of the possible max score). We apply principal component analysis (PCA) to reduce the 768-dimensional embeddings to 2 components on a sample of our reviews.

B. Hierarchical Clustering

To further categorize some of our data, we thought it would be interesting to explore trends within specific movie franchises in the MCU. To do this we employed hierarchical clustering to find which movie franchise each movie best belonged.

In broader research where the focus is not only on MCU movies, comparing only movie revenue could be misleading when assessing a movie's success criteria. The revenue of a successful documentary can be vastly different from that of a successful adventure movie [23b]. However, since all of the movies in our scope fall into the same genre, we should obtain a sort of revenue normalisation.

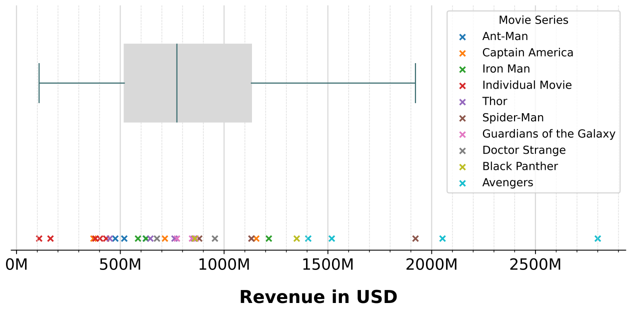

Fig. 1. Box-plot of the *MCU* movie revenues showing a large variety even though they are within the same genre. The categorization of the movie franchise is based on manual labelling.

However, we still see a big difference in the *MCU*'s revenue variance. **(Figure 1)** The above visualization also hints that revenues are somewhat correlated within movie franchises. This tells a story of how sequels might attract the same crowds and carry on a certain momentum in revenue from the previous movie.

But following this logic, how can we then explain the revenue development from prequels to sequels? One of our hypotheses is that the similarity in cast affects how likely one is to watch a sequel, given that they have watched the prequel. If we wanted to test this hypothesis, we would currently not be able to since we could not acquire any data on what movie franchise the movie falls in, if any. For our scope, we could simply fix this by manually labelling, but it would be tedious and complex to apply to bigger datasets. Some movies might explicitly indicate their respective franchise in their title, but the first *Spider-Man* franchise might not have any significant revenue correlation to the later franchise. Ultimately, it would be up to the labelling entity's knowledge, resources and personal opinion to decide, potentially yielding inaccurate labels.

Taking the *Spider-Man* example, one significant indicator is that the cast changes drastically between each franchise. So, if we could tell the difference between each movie's cast, we could use this metric to find clusters of movies and categorize them as being in the same franchise. We could even utilize the release date to tell the order of the movies within the franchise.

1) Cast Distance: But how can we measure this similarity between casts? Certainly, changing out the main character is not the same as changing out one of the extras from each movie. However, *IMDb* states that their ranking strategy of cast order can vary between movies. Ultimately, it is a complex question; the best we can do is make a qualified guess. Here, we found that giving the top 30 actors the same weight yielded the best results for clustering. To represent each movie in our dataset, we employed a 613-dimensional binary vector. Each dimension of this vector corresponds to a distinct actor featured among the top 30 credited cast members of an *MCU* movie. For a particular movie, we constructed this binary vector by indicating whether or not each of the 613 actors appeared in the movie.

2) The Curse of Dimensionality: One should consider *The Curse of Dimensionality* when measuring distances in a high dimensional space. As we increase the dimensions of our vectors, more entropy is introduced, making it increasingly harder to distinguish points from each other. Another major concern is that the computational complexity increases drastically when adding more dimensions. In our case, we didn't experience any major issues given our sample size, but since we are trying to propose a generic solution, we still need to address this major issue. A common solution is to reduce the number of dimensions using principal component analysis (PCA). Here, instead of describing each movie in the real 613-dimensional space, we could possibly describe the cast in a much smaller latent space. This would, of course, introduce some amount of error but make it significantly easier to cluster the movies. In practice, applying PCA here was a bit out of scope for our project, so we decided to improve the method by using the Hamming distance when clustering. This distance metric addresses the issue of increased entropy by only looking at the categorical difference between pairs of data points. In effect, using this metric would only yield a distance of 1 if the only difference between 2 movies was the presence of a given actor.

3) Finding Max Distance: Now that we had established a distance metric, we applied hierarchical clustering to the movies. Because of our limited scope, we could visually inspect the clusters obtained via a dendrogram **(Figure 3)** and guess an appropriate threshold for max distance. However, since we are proposing a generic solution, our approach was to first manually label the movie franchise and then find the lowest threshold value for when the true number of clusters was obtained. This resulted in a max distance metric that we could try to use on bigger datasets if normalised to the number of dimensions. One major concern here, however, is that this normalised max distance metric might work for the *Marvel Cinematic Universe*, but wouldn't be applicable in other types of movies given the possible variation in casting strategies used for making movie sequels across all movies. One possible way to overcome this issue would be to manually sample movie franchises across the landscape and then apply our approach to develop a more generalized metric.

C. Poster Analysis

Another avenue of analysis explored was aspects of the posters used in advertising the movies. We wanted to see if there were any correlations to be found between features of movie posters and, e.g., the revenue of the movies. **Figure 2** shows a sample of our exploratory analysis of extracting color features from a poster, such as average color and the most dominant colors. Other features, such as contrast, were also considered, but after initial plots with these features, no clear trends were found, so further exploration into this analysis ended.

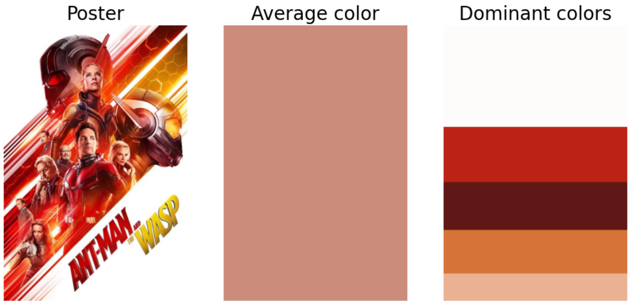

Fig. 2. Color analysis of *Marvel*'s *Ant-Man and the Wasp* Poster. The Left is the raw poster, the middle is the average colour value of all pixels, and the right shows the top 5 most dominant colours in the poster from most to least.

D. Time Series Analysis

The release date of movies and the associated reviews collected can be expressed as a time series, allowing for insights into the evolution of certain variables.

1) Budget and Revenue over time: At first, we conducted an exploratory analysis assessing the yearly budget, revenue and average critic's score. The metadata was grouped by year and added to get the total values of the variables. Later, The Critic reviews per year were averaged and extracted. However, as mentioned before, these scores are binary on whether the review was ”Rotten” or not. This was tackled by obtaining the ratio of not rotten reviews and multiplying them by 10.

2) Movie's perception over time: A main interest during the project was to check how the user reviews of a movie change over time and their relationship to the critic reviews. This was done by extracting the user reviews, grouping them by day, averaging them, and plotting them as time series data. On the other hand, the critic reviews don't span a long time between them, just a month after the release, so the average Critics' Score was computed as before and used as a constant.

It was also agreed to limit the analysis to just one year, as all the movies' lifetimes are quite varied, and the vast majority of *MCU* movies in our Database have at least a year of user reviews.

The Data had to be smoothed via a Rolling window to understand the time series' behaviour better. Here, the Score's cumulative daily average was used to check how the reviews stabilised over time and the cumulative daily variance to understand the development of polarization.

IV. Results

A. Comparing Sequel Revenue with Cast Similarity

To compare a movie sequel to its prequel, we first had to fill in the missing data of what movies were in the same franchise.

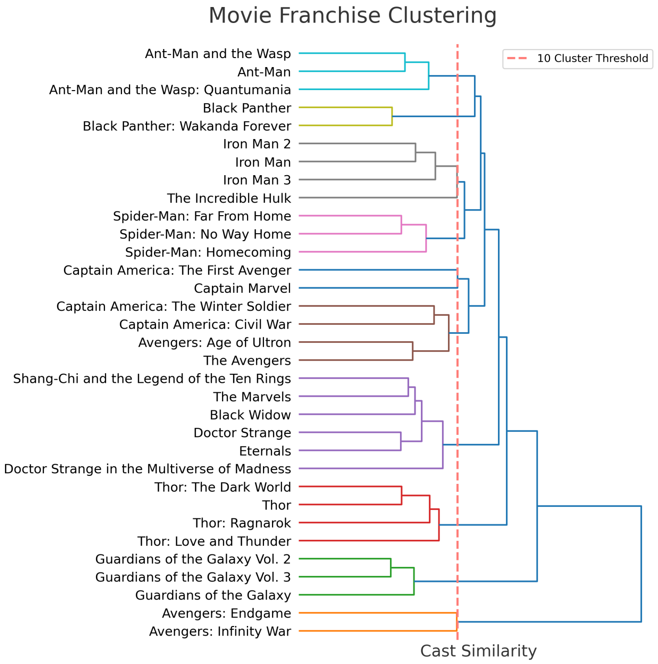

Fig. 3. Dendrogram showing the ten movie franchise clusters obtained by applying hierarchical clustering based on the Hamming distance between the top 30 casts of all *MCU* movies in our dataset. Each leaf node represents a movie, and the distance between any two leaves represents the dissimilarity between their casts. Each distinct line colour on the left of the threshold indicates the respective franchise cluster. The diagram showcases the feasibility of our clustering approach with an 85% accuracy.

The dendrogram in **Figure 3** shows the results of applying hierarchical clustering on movie casts. Of all the 33 movies, only five were misplaced in the wrong franchise, resulting in an accuracy of around 0.85.

Now that we have proposed a solution to obtain the movie franchise data, we would like to explore the within-movie-franchise relationships between revenue and cast similarity. However, for the focus of this analysis, we chose to use our own manually labelled movie franchise data so we do not extend on the small inaccuracy obtained by our clustering approach.

Figure 4 presents the trajectory of all *MCU* sequels, excluding standalone movies. An evident trend emerges: sequels consistently exhibit higher box office earnings than their respective prequels. This observation aligns with data from a recent blog post on datawrapper[23a]. Julian Freyberg, a developer and movie enthusiast, compiled a compelling

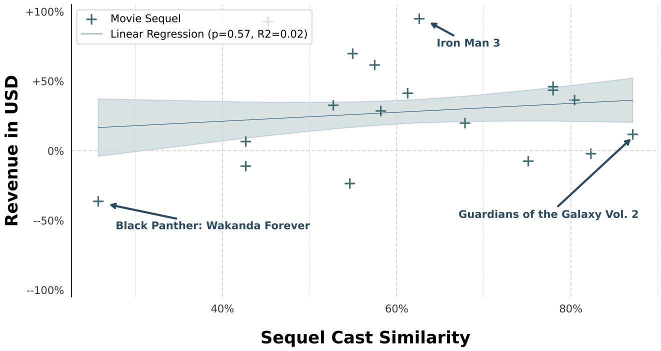

Fig. 4. Scatter plot indicating a weak linear relationship between cast similarity and revenue of movie sequels compared to their respective prequels. The similarity metric is different from the one used in our Hierarchical clustering since we are considering actors earlier in the cast as more important than later ones.

visualization titled *A trilogy is (usually) enough*, illustrating a declining trend in average movie reviews with each sequel. This narrative highlights the inherent trade-off between commercial success and critical appreciation in the world of sequels. While sequels often reap greater financial rewards, they may struggle to replicate the initial excitement and novelty that fueled the prequel's success.

While creating the visualization in **Figure 4**, we realized that the distance metric employed in hierarchical clustering was inadequate for portraying the intended narrative. Our aim was to depict the story of how the movie *Black Panther: Wakanda Forever* fell short of its predecessor's box office success, a consequence of Chadwick Boseman's untimely passing, which left a void in the titular role. To capture the importance of the actors, we found that the equation

$$importance = 1/i$$

where i is the index that the actor appeared in the cast, yielded the best results. Using the hyperbolic formula, we capture how changing out the main character, like with the *Black Panther* example, has a way higher impact than actors appearing later in the cast. Additionally, as we believe actors disappearing or being demoted have the highest impact on the success of a sequel, we didn't want to punish actors advancing in the cast ranking when measuring the distance between casts. Given the high p-value of 0.54, we could not reject our null hypothesis even after trying to adapt the distance metric. However, the visualization still added valuable information to the revenue correlation of sequels concerning their prequels.

B. Revenue Correlation

When analyzing a movie's overall success, a simplified way to measure it is to look at its box-office revenue. Here, we tested the Pearson correlation between all recorded features and the revenue to measure which features have the strongest correlation.

1) Maximizing Revenue: In our analysis, **Table IV** shows that the strongest linear relationship we found was between the reported budget of the movie, the percentage of reviews containing perfect scores, and the mean critic score of the movie being the best indicators of revenue. We also found that the correlation between the average critic rating was stronger than the average audience score.

Variable	Pearson-Correlation	P-Value	R-Square
Budget	0.69	< 0.001	0.477
% of Scores = 10/10	0.499	0.003	0.249
% of Scores = 9/10	0.241	0.183	0.029
% of Scores >= 9/10	0.429	0.0143	0.184
Mean Critic Score	0.419	0.016	0.176
Mean Audience Score	0.279	0.122	0.078

TABLE IV
CORRELATIONS AND P-VALUES FOR SELECT VARIABLES, WHEN USING THE *Variable* COLUMNS AS X-AXIS AND REVENUE AS Y-AXIS.

With a range of correlations between $[0.27 : 0.69]$, we can observe a somewhat strong increasing linear relationship between the tracked variables and revenue. With a critical value $\alpha = 0.05$, we can check the significance of the tracked variables with a p-value $P < 0.05$ and confirm that the results are statistically significant (for select variables). This does indicate that the highest-earning movies have a large fraction of people rating them close to perfect. We also note that the highest-earning movies have higher critic scores than audience scores. We can use this as motivation for analysing the written critic reviews in our review embedding analysis.

2) Maximizing Profit: Another perspective on what a successful movie would be is assessing a movie's ability to make back its budget. We compared the collected features to the profit ratio: $\frac{Revenue}{Budget}$ **(Figure 5)**. We deem the results of this analysis statistically significant, which we can confirm with our critical value of $P < 0.05$ and note a positive linear relationship.

C. Embeddings

After establishing that the average critic score gives us an overall clearer picture of a movie's earnings, we explore further the review embeddings created in the Methods section. Using embedding models to classify reviews, we find that by using a 2-cluster K-Means clustering algorithm, we can predict with an accuracy of 0.84 whether the review contains a "good" or "bad" score. In this example, we only use reviews written on *Rotten-Tomatoes* by verified critics since our model best separates them linearly. **(Figure 6)**

D. On analyzing time series data

Figure 7 shows us that, except in 2009 and 2020, at least one *MCU* related movie has been released yearly. The latter is due to the Global pandemic. The number of movies released yearly also stayed relatively consistent, with two or three in most years. Despite this, each year's budget has doubled since the beginning, meaning the production has become increasingly expensive. This investment has given good results, as we can appreciate when reviewing the yearly

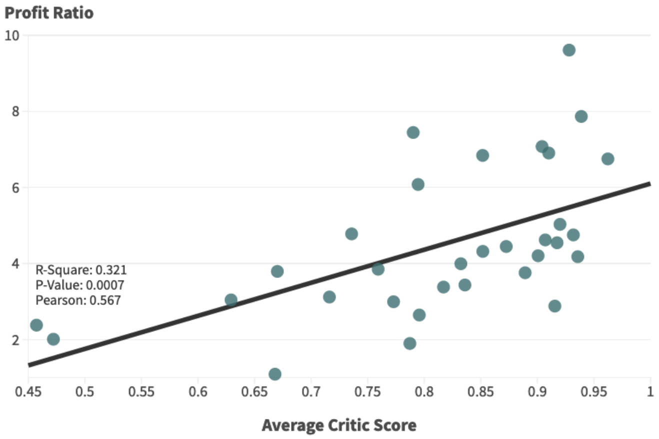

Fig. 5. Correlation of average critic score and the ratio of how many times a movie made back its budget in revenue. X-axis denotes the average critic score ($\frac{\#PositiveReviews}{\#TotalReviews}$). Y-axis denotes the *profit ratio* ($\frac{Revenue}{Budget}$)

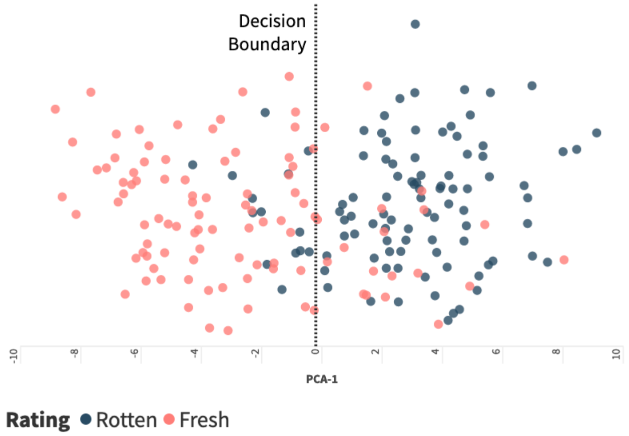

Fig. 6. Scatter plot of PCA-1 (X-axis) and PCA-2 (Y-axis), colored by the rating assigned by the author of each review. The highlighted decision boundary yields a 0.84 accuracy on classifying Rotten/Fresh Reviews. The decision boundary is constructed by an unsupervised k-means clustering model, initialized with 2 clusters.

revenue. It also shows a sharp increase until reaching its peak in 2019 (the year *Avengers: Endgame* released). After this point, the revenues decrease even with the same budgets and number of movies released. Conversely, the average score is not affected by the budget or revenue in any meaningful way.

Given the noise of our initial plots, we decided to pivot towards a case study. Here, three movies were selected to analyze. The movies were chosen according to our analysis of relative franchise success in terms of revenue. The movies of interest are annotated in **Figure 4**: *Iron Man 3* as a successful example. *Guardians of the Galaxy: Volume 2* as a mediocre success. And *Black Panther: Wakanda Forever* as an example of a relative financial failure. **Figure 8** shows the experiment regarding audience perception over time. In the analysis itself,

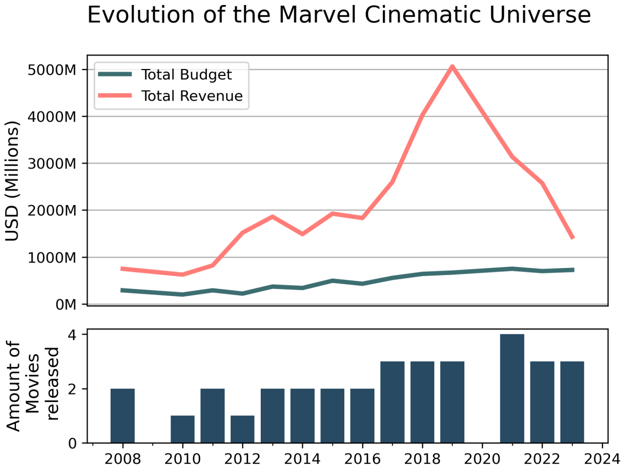

Fig. 7. Time series plot showing the evolution of *MCU* movie statistics. Here, we generally see how budget and revenue go up over time. However, we also notice a big shift in revenue since the 2020 pandemic, and the most recent movies possibly not reaching their full revenue potential yet.

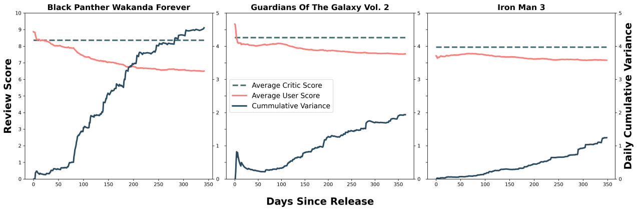

Fig. 8. Time series plot of cumulative user review score and cumulative variance of our three test cases over their first year of release. The average critics score is also plotted for reference.

a few things from the plots stand out: The first one is the slope in the averages, where all the movies experience varying degrees of decline in their Average score as time progresses. We also note that the more successful a movie is, the closer it stays to the critics' average over time.

When assessing the variance, we note that with an unsuccessful movie like *Black Panther: Wakanda forever*, we can see that the variances increase a lot over time. In this case, it reaches a variance of 4. Meanwhile, movies we deemed successful have a smaller increase in their variance, with *Iron Man 3* barely reaching a variance of 1.

V. Discussion

After analyzing our dataset, we have obtained a diverse range of findings. We will now try to understand and rationalize these results to the best of our abilities.

A. Review Scores

In our analysis, we find that the review-based variable with the strongest correlation to profit is the percentage of 10/10 review scores. We also note that verified critic scores have the second-highest correlation. While we could find moderate to strong correlations, we have little say in whether this is a correlational effect or a causation. From the results, we do have some hypotheses/assumptions as to why.

1) Perfect Reviews: The results showed that a clear indicator of box-office revenue is a large fraction of audiences, giving a perfect 10/10 score. We use this result to argue for an overall tendency for audiences to watch a movie if people don't just like it but love it. We argue that audiences are drawn to movies which bring out more visceral opinions. This also suggests that in this landscape of movies, only movies which are disproportionally well received by audiences will guarantee a large box-office revenue. This is different from looking at the mean review score, as a movie can have a high mean rating but contain a lot of scores below 10, which the data indicates hurts box-office revenue.

2) Critic Reviews: As noted in **Table IV**, the critic score correlates more with the box-office revenue than the audience score. We used this knowledge as motivation to analyse the nature of the critic reviews further. In the PCA analysis, we found that we can separate the opinions of verified critics with a fairly high accuracy. While this conclusion does not strictly help us evaluate the success of a movie, it does help us understand the nature of our data better. We also use the results of this analysis to prove that our data follows an underlying structure/nature (which the chosen embedding model captures).

B. On Hierarchical Clustering

The hierarchical clustering dendrogram in **Figure 3** indicates that our hierarchical clustering method might be valid for synthesizing the categorization of movies into their franchises. While the 0.85 accuracy shows it is not perfect, it still adds valuable insights into what movies are correlated with each other. Some misplaced movies might indicate that while within the same franchise, some may have a stronger relationship with other movies. This is showcased in the movies *Avengers: Endgame* and *Avengers: Infinity War*. In these two blockbusters, *Marvel* brought together almost all the superheroes from their previous movies, filling up a big part of the top 30 casts with actors not previously credited in the prequels. Additionally, the dendrogram shows that if we were to cluster the *MCU* movies into 2 clusters, it would result in one cluster with the above-mentioned *Avengers* movies and all the other *MCU* movies in a second cluster. Ultimately, our hierarchical clustering approach is limited to only considering the cast similarity. To capture edge case scenarios like the above-mentioned, we would have to extend the approach to encompass more sophisticated inputs.

C. Homogeneity

One of the revelations of this project is that the scope we decided on would inherently invalidate some types of comparative analysis. This also means we could not compare movies with different characteristics like genre and budget. While the similar budget range of our movies ensures that we don't have any outliers with disproportionally high revenue compared to its budget, it also more-or-less meant that the budget would end up being the best estimator for a movie's revenue. This means that we cannot generalize the findings of this project to the wider movie landscape, as we are looking at a small sample.

D. No Direct Financial Failures

None of the movies in our scope were considered financial flops (based on reported box office revenue, excluding marketing expenditures). Consequently, we could not directly compare the characteristics of a financially unsuccessful movie to those of a successful one. Nonetheless, within our project's limitations, we could contrast these movies' relative successes and failures.

E. Life of a movie over time

Our time series analysis indicated that consistently high user ratings are a hallmark of successful movies. Conversely, sequels deemed ”unsuccessful” by our metrics experience an unstable average user score, leading to a higher cumulative daily variance. In contrast, sequels with a positive relative revenue maintain relatively stable review scores, even after an extended period. Consequently, their variance remains relatively low over time.

Returning to our analysis assessing correlations between online reviews and revenue, we find that a high fraction of 10/10 scores yields more revenue. By combining the findings with the time series analysis, we argue that these findings show that financially successful movies tend to have a high fraction of 10/10 scores and a common consensus that does not change drastically over time.

This, of course, is an incomplete assessment, as we are only looking at movies with a generally positive reception due to the scope of the project previously discussed. It's hard to believe, and against common sense, that a movie that's ”consistently bad” critically wise would be a successful one, but there have been few and sparse cases where movies considered bad during its initial reception become ”cult classics” and regain some revenue due to retroactive success-factors like DVD sales.

F. Types of Audiences and when do they watch

Regarding more discoveries in the Time Series, the constant negative slope in the movies' cumulative average could be attributed to the different audiences who consume these movies: The ”Fandoms” and the ”General Audience”. The more passionate audience around the *MCU* and Superhero movies, in general, will certainly watch the movie first and probably give a higher, likely skewed, review score due to their interests. It's even possible. After this initial uplift in user reviews, the more ”regular” audience watches the movie, and as they're not as invested, their reviews might be as biased. These reviews, of course, reduce the cumulative average to something that represents the movie better.

G. Future Work

Given the current project's scope, we propose exploring certain avenues in greater depth.

1) Moving Beyond Contextual Data: One of the main weaknesses of the project is our exclusive focus on the contextual data of the movies. In future work, we recommend looking deeper into the screenplay data if possible. For example, one could transcribe all the movies in the dataset and perform semantic analyses from the scripts. Another angle could be conducting statistics of how many different sceneries were present based on image analysis of the frames over time. However, we understand how this, in most cases, might be an impossible task since the movie data is considered intellectual property, and this kind of processing would be illegal.

2) Extra spending and revenue sources: To get the full picture of a movie's success, it's not enough to take the raw budget and revenue from ticket sales. Much more capital is invested through Marketing campaigns and Publicity on different platforms. Additionally, most movies probably have more than one revenue source, such as DVD sales and streaming income. Investigating how much marketing campaigns affect users' and critics' perceptions of a given movie could be interesting. Furthermore, it is imperative to check if the movie rebounded its investment by other means and how long it took. Here, we could get a closer look into what movies were actually financially successful.

3) Poster Analysis: While the explored poster analysis proved not useful for our particular analysis, it could be that it is useful in different research in conjunction with other data. One use case of this could be identifying how different visuals attract certain parts of the population and adapting marketing strategies accordingly.

VI. Conclusion

We initially set out to uncover the key contributing elements to a movie's success. We gathered, processed and analysed movie metadata and online reviews and discovered statistically significant patterns in our data. Here, we found that a movie's budget has the strongest correlation with its revenue.

In the process, however, it became clear that success is relative and not just defined by total revenue or average user reviews. Limiting our scope to a successful movie studio provided a somewhat homogeneous dataset with no direct financial failures. However, we discovered that, on average, a *Marvel* sequel makes more money than its prequel, indicating one way of defining the success of franchises within the *Marvel Cinematic Universe*. Even then, a movie can still fail in the public perception. Here, we found a strong correlation between the percentage of reviews with a 10/10 score and the total revenue. Comparing this correlation with the mean audience score indicated that maximizing the percentage of admirers yielded more income than having people accept it on average. Additionally, one case study on the daily average variance showed how a more stable consensus is desirable over time.

Even with these strong indicators, we must still acknowledge the underlying challenge of finding patterns indicating success in contextual data of the actual data: The screenplay. Here, a mixture of acting, direction, budgeting, etc., yields a complex video data format we cannot access or process. Consequently, one should strive to require and consider as much contextual data about the movie as possible. However, due to the limitations of the publicly available contextual data and the time constraints of our project, we made educated guesses on a list of focus points. Here, we searched for revenue-contributing patterns in review embeddings, movie posters, and sequel cast similarity without finding any statistical significance. However, in the process, we still confirmed the relevance of the language of the reviews by showing its correlation with the score. Additionally, we built a useful dataset of movie posters and proposed a generic solution to cluster movie franchises based on cast similarity using hierarchical clustering.

While these findings may apply to our specific project, they do not necessarily guarantee their validity in other domains. Nevertheless, they highlight the remarkable ability of data wrangling and exploration to unveil unexpected patterns. We've observed positive outcomes in our context, but this doesn't always hold. This underscores the importance of exercising caution when dealing with data in the wild, as we cannot always foresee the unforeseen consequences that may arise.

References

[23a] *Is there a movie franchise crisis?* https://blog.datawrapper.de/is-there-a-movie-franchise-crisis/. 2023 (cit. on p. 4).

[23b] *Most popular movie genres in the United States and Canada between 1995 to 2023, by total box office revenue.* https://www.statista.com/statistics/188658/movie-genres-in-north-america-by-box-office-revenue-since-1995/. 2023 (cit. on p. 3).

[23c] *Rizz-Kings GitHub Repository.* https://github.com/rizz-kings/data-in-the-wild. 2023 (cit. on p. 2).

[23d] *The Movie Database.* https://www.themoviedb.org/. 2023 (cit. on p. 1).

[23e] *tmdbsimple: A wrapper for The Movie Database API v3.* https://github.com/celiao/tmdbsimple. 2023 (cit. on p. 1).

[23f] *Top 1000 Websites By Ranking Keywords.* https://dataforseo.com/free-seo-stats/top-1000-websites. 2023 (cit. on p. 1).

[RG19] Nils Reimers and Iryna Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov. 2019. URL: http://arxiv.org/abs/1908.10084 (cit. on p. 2).