CS414

2022-2023 Spring

Project Report

Beyond the Budget Dive into Marvel Universe

Serra Sadır 28201

Irmak Özügüzel 28258

Emre Argın 28019

Hakan Körpe 26397

Cengiz Han Ermiş 28378

ABSTRACT

The project aims to understand the factors influencing the success of Marvel movies/series by examining their box office earnings and exploring relationships between movies using a graph-based approach. The analysis begins by processing movie data, including information on movie names, box office earnings, and actors. A graph is constructed, with movies represented as nodes and their box office earnings determining the node sizes. The graph is enriched by adding edges between movies that share a significant number of actors, indicating potential connections or similarities between them. Additionally, separate graphs are created using release date and critic scores as attributes, allowing for the visualization and exploration of temporal trends and critical reception. To further investigate the relationship between movie success and financial aspects, a regression analysis is conducted. The budget is considered as the independent variable, representing a potential predictor of box office earnings. By fitting an ordinary least squares (OLS) model, the project examines the relationship between budget and box office earnings. The regression results provide valuable insights, including the regression coefficient, p-value, R-squared, and F-statistic, indicating the strength and significance of the relationship between the two variables. By combining graph analysis and regression techniques, this project provides a comprehensive understanding of movie success. The graph visualization allows for a visual exploration of movie connections based on shared actors, while the regression analysis uncovers the influence of budget on box office earnings. The findings from this project could contribute to the understanding of factors driving movie success and inform decision-making in the film industry.

Introduction

1.1 Background and Motivation

Superhero movies, especially those from Marvel Comics, have been making a splash in the film world for the past decade. But what makes these films successful? It's not just the Marvel name. Things like star power, big budgets, when the movie comes out, and good reviews also seem to matter. Still, how all these factors mix together to make a hit film isn't very clear.

In our group project, we're trying to understand this mix better. Using data analysis and network analysis, we're looking for patterns and connections between all these factors. We chose Marvel films for our study because they give us lots of data to work with.

Our goal is to make sense of what makes a film successful. We hope this will help students who are interested in this topic, people working in the film industry, and academics get a better understanding of the movie business. By figuring out how different factors work together to make a film successful, we think we can provide useful information for people making and marketing movies.

1.2 Research Question and Objectives

Against this backdrop, the primary research question for this study is: Does the budget of Marvel movies significantly impact their success, as measured by box office revenue and critical acclaim?

To address this question, the study aims to achieve the following objectives:

Examine the relationship between movie budgets and success metrics, such as box office revenue, and critical acclaim, in the context of Marvel movies.

Utilize network science methodologies to construct and visualize a network graph, incorporating variables such as release date, duration, starring, and budget, to explore the complex interplay between these factors and success metrics.

Analyze and interpret the network graphs to identify patterns, trends, and potential correlations between budget and success metrics.

Compare and contrast Marvel movies with different budget levels to determine if budget is a significant differentiating factor in terms of success.

Provide insights and implications for filmmakers, studios, and researchers in the film industry regarding the relative importance of budget in achieving movie success.

1.3 Significance of the Study

This study matters for several reasons. One of the key groups that can benefit is the film industry. Filmmakers, producers, and marketers can gain insights from the study to understand what contributes to a movie's success. These insights can inform decisions about how much money to spend on a movie, who to cast, when to release a movie, and how to advertise it.

The study also has implications for academic research. It adds to our understanding of movie success and demonstrates how we can use data science to shed light on the film industry. The approach used in this study could serve as a model for similar research on other types of films.

Finally, moviegoers stand to benefit from this research as well. Knowing what makes a movie successful can help audiences make better choices about what to watch. It can also lead to more informed discussions about movies.

In a nutshell, this study illustrates how data can be used to dissect complex systems like the movie industry. Its findings can help various stakeholders, from filmmakers to audiences, navigate the world of cinema more effectively.

2. Literature review

2.1 Theoretical Framework

The theoretical groundwork for this study focuses primarily on the principles of network science and film industry economics. These concepts equip us with a robust method to explore the intricate network of relationships within the Marvel Cinematic Universe. Network science offers a structured methodology for examining the interconnected attributes of the MCU films, such as the cast, box office earnings, budget, release date, and critic scores. These elements provide the foundation for understanding the commercial success and financial performance of these films.

2.2 Previous Studies on Movie Success Factors

The link between movie success and its various determinants has been explored in several studies. Tomaric proposed an argument for the potential of low-budget films to achieve success. This study emphasizes the importance of resourcefulness and effective filmmaking practices, suggesting that financial constraints need not limit the potential for professional and successful filmmaking.

Lash and Zhao presented a framework for predicting movie success, exploring a variety of factors including the cast, content, and release timing. Their study offered a more holistic approach to understanding movie profitability, highlighting the interconnectedness of various movie attributes and their collective impact on the film's performance.

Another study by Jonas Sebastian Krauss focused on predicting movie success and Academy Awards using sentiment and social network analysis, demonstrating the predictive power of these modern analytical techniques. By evaluating public sentiment and the social networks of movie industry professionals, this study offered new ways of anticipating movie success.

The impact of star power on financial success was the focus of the study "Movie Stars and the Distribution of Financially Successful Films in the Motion Picture Industry." John Sedgwick, Micheal Pokorny. This research emphasized the value of high-profile actors in enhancing a movie's financial performance, thereby reinforcing the industry's common practice of casting established stars.

Finally, "The Determinants of Motion Picture Box Office Performance: Evidence from Movies Produced in Italy" by Bagella, Michele, and Leonardo Becchetti. conducted a

multifactorial analysis of box office performance. Despite being specific to the Italian film industry, this study's comprehensive approach enriched the understanding of movie success factors by considering a wide range of influences from star power to genre, seasonality, and more. Together, these studies paint a complex picture of movie success, emphasizing the interplay of various factors from financial resources and star power to sentiment analysis and timing.

2.3 Limitations and Gaps in Existing Research

While numerous studies have contributed significantly to understanding movie success factors, certain limitations and gaps are evident when the research is focused on unique cinematic universes like the Marvel Cinematic Universe (MCU).

Most studies look at one success factor at a time, like budgets, star power, or release dates. However, in reality, these factors don’t exist independently, especially in a complex universe like the MCU where everything is connected. Existing research often overlooks this interplay. Our research will fill this gap by exploring the interconnected nature of MCU, and how and what makes a movie successful by creating multiple network models.

3. Methodology

3.1 Data Collection and Preprocessing

In this study, we wanted to understand what makes a film successful. We needed a lot of different data for this, like objective facts and figures as well as subjective ratings from viewers and critics. We found all of this information on Wikipedia, a huge online database that's free for anyone to use.

We used tools in the Python programming language, specifically the BeautifulSoup and requests libraries, to automatically collect this data. BeautifulSoup helped us understand the structure of each Wikipedia page, and Python's regular expressions (regex) let us pinpoint and take out the exact pieces of information we needed. Once we had all the data, we organized it into a pandas DataFrame, a kind of data structure in Python that's easy to work with and analyze.

The initial data points, namely the movie names and the links to their individual Wikipedia pages, were sourced directly from a comprehensive list of films based on Marvel Comics publications. These links served a two-fold purpose; they provided a reliable reference to the original data source and facilitated the extraction of additional detailed information from

each movie's individual Wikipedia page. This information included the main actors or 'Starrings', the box office gross, the production budget, and the film's running time.

Additionally, data points such as the release date and the critic score were gathered from another table on the original list page. To streamline our dataset and focus on relevant details, release dates were stored as years, as the precise day and month of release were not critical to our analysis. Critic scores were procured from Rotten Tomatoes and provided on a scale of 0 to 100, offering a standardized measure of critical reception.

The specific data points collected were chosen carefully due to their significance in the film industry:

1. **Movie Name:** This serves as a unique identifier in the dataset, enabling cross-referencing with other sources if needed.
2. **Link:** The URL to the film's Wikipedia page was saved for reference, and to facilitate extraction of further information if required.
3. **Starrings:** The film's cast was included as star power can significantly influence box office performance.
4. **Box office (in Million Dollars):** Box office gross acts as a direct indicator of a film's financial success, allowing us to measure overall popularity.
5. **Budget (in Million Dollars):** Knowing the production budget can give insights into the scale of the film, which could have correlations with box office success.
6. **Time (in minutes):** The film's running time may impact audience preference, and thus, box office results.
7. **Release Date:** The release date can affect a film's success due to seasonality, competition, and other market factors.
8. **Critic Score (from Rotten Tomatoes):** A measure of critical reception, this score can influence audience viewing decisions, and thus, the film's financial success.

By merging both methodical data extraction techniques and the careful selection of significant film characteristics, this structured data frame allows for a comprehensive analysis of the factors contributing to a film's success.

3.2 Network Construction

The network was constructed using Python to map out connections between various movies. Each movie was denoted as a node within this network, the size of which was determined by its box office revenue scaled in relation to the film's budget. The edges, or the

lines connecting nodes, signified shared characteristics between movies. These shared elements could be common actors, movie length, release year, critic ratings, or financial success at the box office. The strength of the connections was represented by the edge weights, which were based on the level of shared characteristics.

3.3 Network Visualization

The visual representation of the network was facilitated by a 'spring layout'. This layout ensures the optimal distribution of nodes, offering an easy-to-understand visualization of the relationships between movies. The nodes were differentiated by size, with larger nodes indicating movies that yielded higher box office returns in relation to their budget. The edges, representing shared characteristics, provided an intuitive visual understanding of the links between different movies.

3.4 Network Analysis Metrics

The network analysis involved the evaluation of node sizes and edge weights. Node sizes, depicting box office success in relation to the budget, were quantitatively analyzed to determine the financial success of the movies. Edge weights, on the other hand, provided a metric for evaluating shared characteristics between movies. By examining these weights, the level of commonality could be quantitatively understood, providing valuable insights into the links between different movies.

3.5 Statistical Analysis Methods

To augment the network analysis, a regression analysis was performed, exploring the predictive relationship between a movie's budget and its box office success. This statistical method facilitated the determination of patterns in the data and the likelihood of a correlation between budget and financial success. Subsequent tests were carried out to ensure the statistical significance of the results, thereby validating whether the observed patterns were non-random occurrences.

4. Results and Analysis

4.1 Descriptive Analysis of Marvel Movies Dataset

In this section, we present a descriptive analysis of the Marvel movies dataset. The dataset includes information on various attributes such as budget, box office performance, critical reception, starring actors, duration, and release dates. We provide an overview of these variables and their distributions to gain a comprehensive understanding of the data.

4.2 Visualization and Interpretation of Network Graphs

We constructed network graphs to explore the relationships between Marvel movies based on similarities in budgets, starring actors, release dates and durations. These graphs visually represent the connections between movies, highlighting clusters and influential nodes. Through the analysis of centrality measures, including degree centrality, closeness centrality, and betweenness centrality, we identify movies that play pivotal roles in the network. By visualizing and interpreting the network graphs, we observe that certain movies, such as "The Avengers" and "Guardians of the Galaxy," serve as important hubs within the Marvel movie universe.

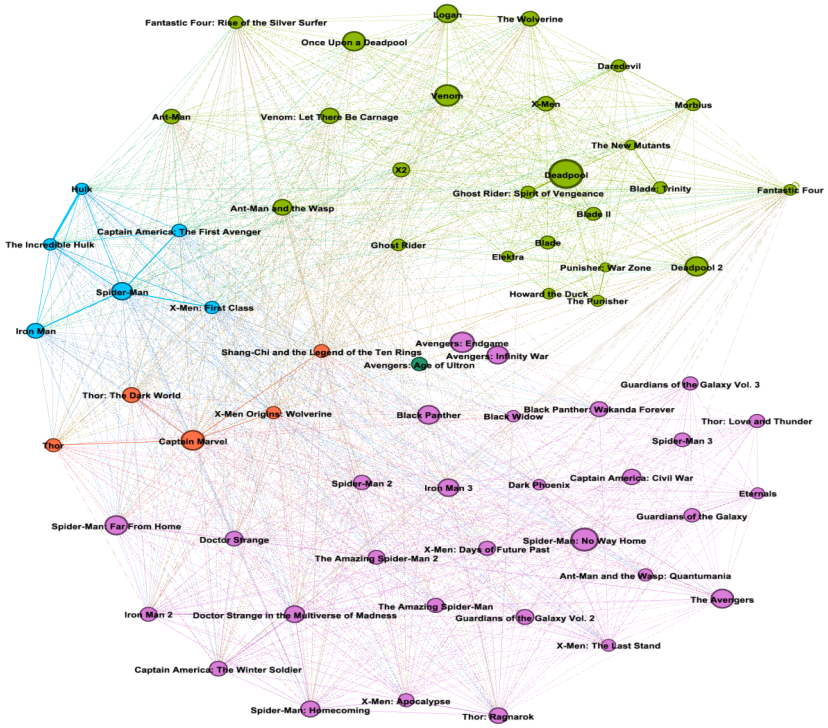

Fig 1. Nodes are movies - Edges are budgets (diff < 80) - Node sizes are success rate

In Fig 1 we aim to see the budget and success relation. The graph is divided into 4 sections in terms of budget levels. The green section has relatively higher success rates than other sections whereas there are also remarkable unsuccessful movies and also there are successful movies in the pink section. Thus, it is possible to say that although the budget has a significant effect on movie success, it cannot be the only variable. Analyzing the graph, we can observe the following:

There are several nodes with high closeness centrality, indicating that they are well connected to other nodes in the graph. These nodes include Spider-Man, Hulk, Fantastic Four, Iron Man, Doctor Strange, Captain Marvel, and Spider-Man: Far From Home.

The betweenness centrality values for most nodes are relatively low, indicating that they do not play a significant role in connecting other nodes in the graph.

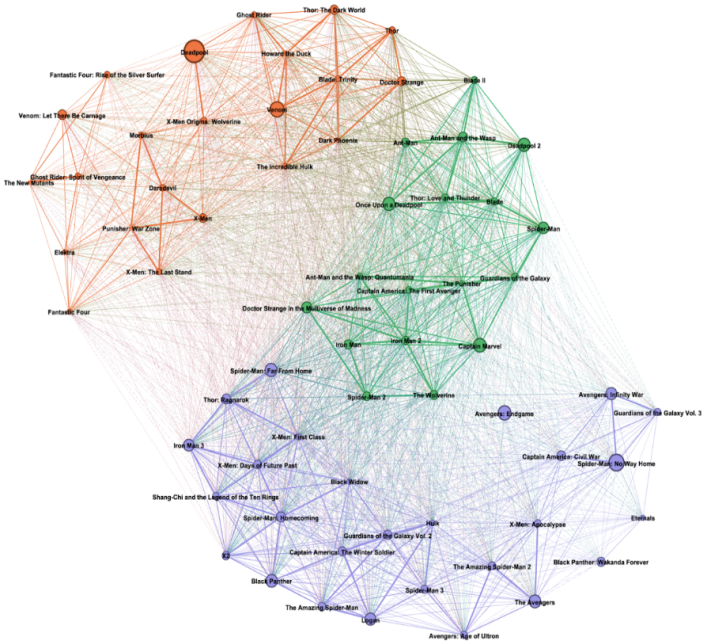

Fig 2. Nodes are movies - Edges are movie duration difference (diff < 40) - Node sizes are success rate

Some nodes have lower success rates (smaller node sizes) compared to others. This could indicate that these movies had lower box office returns or lower critic scores relative to their budgets. Some examples of movies with lower success rates include Howard the Duck, Blade, The Punisher, and Elektra. On the other hand, there are nodes with higher success rates (larger node sizes), indicating that these movies had higher box office returns or higher critic scores relative to their budgets.

Examples of movies with higher success rates include Spider-Man, Hulk, Iron Man, Doctor Strange, and Captain Marvel.

The regression analysis shows a positive relationship between the budget of a movie and its box office returns. The coefficient of the budget variable is 4.735, indicating that, on average, a $1 million increase in the budget is

associated with a $4.735 million increase in box office returns. The p-value is very low (1.913e-12), indicating that this relationship is statistically significant.

In Fig2. it is aimed to see the relationship between the duration and the success of the movies. The graph is divided into 3 sections in terms of the time difference. (The same color represents a similar time difference.) According to the p-value of this network the length of the movie is a significant factor that affects the box office earnings with other variables. From the positive regression metric every minute added to a movie's runtime, The Box office earnings

increase by approximately 19.73 million dollars on average holding other variable constants. To sum up, this model suggests that longer movies tend to have higher Box Office earnings.

Visualizing the Influence of Release Dates on Box Office Performance:

Fig 3 Nodes are movies - Edges are movie release date difference (diff < 3) - Node sizes are success rate

date_threshold = 3:

Network Construction and Edge Weighting:

In the code, the graph represents the network of movies, where each movie is a node and the connections between them, or edges, denote the relatedness based on their release dates (Fig3). The size of each node, representing a movie, is calculated using the film's critic score, box office revenue, and budget. This measure provides an estimate of the relative success of the movie, with larger nodes signifying more successful films. The edges between nodes are weighted based

Overall Connectivity:

The graph shows a significant level of connectivity, with many movies having moderate to high centrality measures and clustering coefficients.

The Avengers-related movies (e.g., "The Avengers," "Avengers: Age of Ultron," "Avengers: Infinity War") are central and highly connected within the graph, indicating their significance in the Marvel cinematic universe.

Movies like "X-Men," "Captain America: The Winter Soldier," and "Guardians of the Galaxy" have notable influence and act as important bridges between different parts of the graph.

Evolution of the Marvel Cinematic Universe:

The centrality measures and clustering coefficients suggest that the graph captures the evolution of the Marvel Cinematic Universe, with newer movies being highly connected and influential, like "Avengers: Endgame" and "Spider-Man: No Way Home."

The graph also includes standalone movie series like "Blade," "Spider-Man," and "X-Men," which have their own clusters within the larger Marvel universe.

date_threshold = 5:

Network Analysis:

"Ant-Man and the Wasp" and "Venom" are the most central movies in the network, having the highest closeness centrality, eigenvector centrality, and degree centrality (Fig4). This suggests that these movies are closest to all other movies in the network, have a high influence, and are directly connected to many other movies.

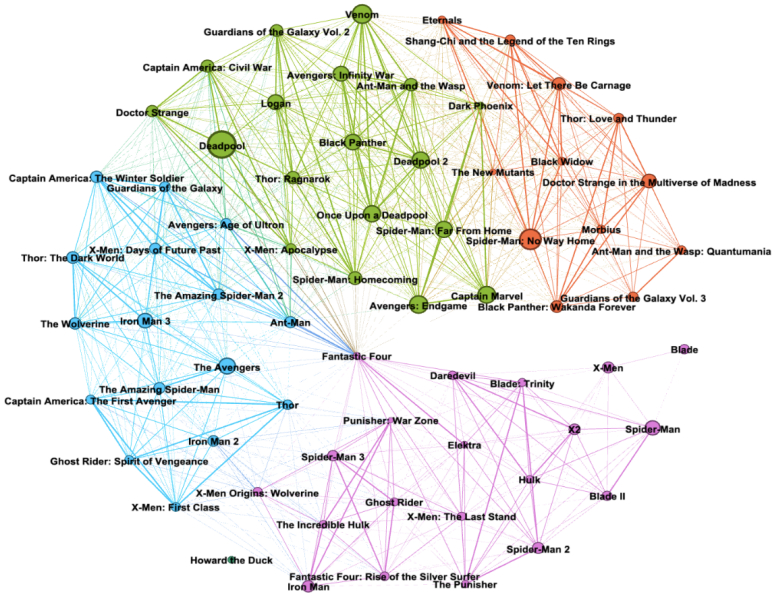

Fig 4 Nodes are movies - Edges are movie release date difference (diff < 5) - Node sizes are success rate

"Guardians of the Galaxy Vol. 3" and "Ant-Man and the Wasp: Quantumania" have the lowest degree of centrality, indicating these movies have the fewest direct connections to other movies.

"Morbius", "Doctor Strange in the Multiverse of Madness", "Thor: Love and Thunder", and

"Black Panther: Wakanda Forever" have high clustering coefficients, suggesting that the movies directly connected to these are also highly interconnected.

Regression Analysis:

An OLS regression was performed to predict the box office results based on the release date of the movies.

The analysis suggests that for each unit increase in the release date, the box office revenue increases by approximately 29.34 million dollars, and this relationship is statistically significant (p-value < 0.05).

However, only about 18.4% of the variation in box office results can be explained by the release date, as indicated by the R-squared value. This suggests that there are other factors not captured in this model that also influence the box office success.

Fig 5. Nodes are movies - Edges are movie release date difference (diff < 7) - Node sizes are success rate

date_threshold = 7:

Analysis for this graph:

The regression coefficient for the release date is 29.34 (Fig 5), indicating that, on average, for each unit increase in the release date, the box office revenue is estimated to increase by approximately 29.34 million dollars.

The p-value associated with the release date coefficient is 0.00026, indicating that the relationship between the release date and box office revenue is statistically significant. The R-squared value of the regression model is 0.184, suggesting that the release date explains about 18.4% of the variation in box office revenue. However, it is important to note that there are other factors not captured in this model that also contribute to the box office success of movies.

Overall, the regression analysis provides evidence of a positive relationship between the release date and box office revenue, but it indicates that other factors beyond the release date are also influential in determining the financial success of movies.

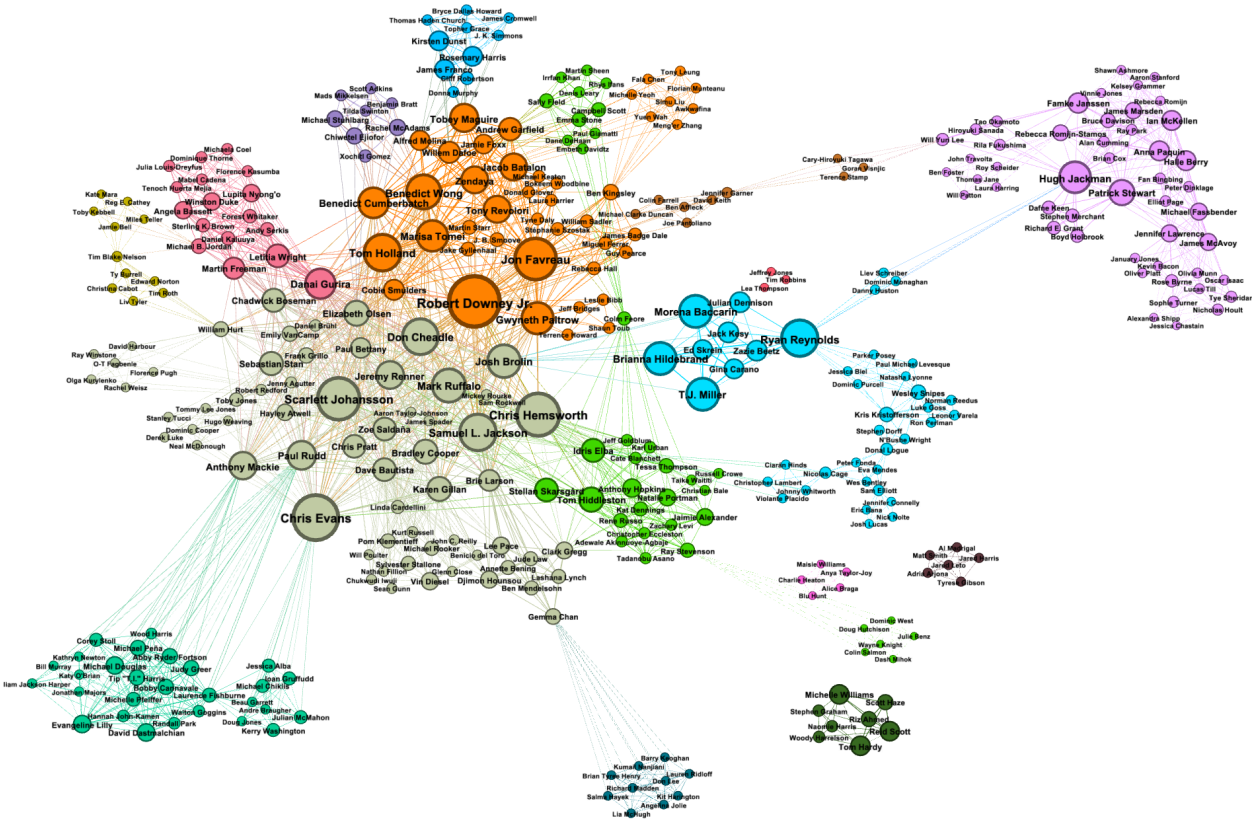

Fig 6 Nodes are actors – Edges are actor collaborations – Node sizes are average movie success rates.

Analysis for this graph:

The network is constructed based on making the nodes as actors in MCU movies, node sizes as the average movie success rate for the actor (total success for the movies that the actor occurred in / number of movies actor occurred in), edges as the collaboration between actors and edge weights are determined by the average success of the movies that the actors collaborated in.

In this actor-collaboration network, we can see that, actors with the most centrality degree are Robert Downey Jr, Chris Evans, Scarlett Johansson, Don Cheadle, and Chris Hemsworth in descending order. This is expected since these actors, in general, had both individual movies and squad movies such as Iron-Man series and Avengers series for Robert Downey Jr. On the other hand, actors with only individual movies or only squad movies tend to fall short on centrality.

For the node size comparisons, like centrality degrees, actors such as Robert Downey Jr., Chris Evans, Chris Hemsworth, Scarlett Johansson, and Jon Favreau have the biggest node size due to their average movie success rates. Since collaborating with huge celebrities and involving in hit series tend to lead to a successful movie occurrence. However, even though the centrality degree for Ryan Reynolds and Hugh Jackman being low, they demonstrate successful occurrences as we can see from their node sizes. Despite the low centrality, Ryan Reynolds is placed at 8^{th} seed for the most successful actor by average movie success rate. Thus, we can infer actors like Ryan Reynolds and Hugh Jackman who do not score high in centrality but have successful occurrences, definitely made an impact on that movie to be successful since they do not collaborate with huge celebrities or took part in squad movies as much as other successful actors. Therefore, their individual effects are much clearer in those movies.

While analyzing the network, we can see some actors having huge successes even if they are not so well known, such as Brianna Hildebrand. This is because she took part in the Deadpool series which is a huge success. Having a low number of movies occurred in and those movies being successful can lead to results like this, but we need to keep in mind that these actor successes are related to well-known series or some exceptions.

To conclude, we can infer that collaboration of huge celebrities such as Robert Downey Jr, Chris Evans, Tom Holland, Chris Hemsworth, Scarlett Johansson etc. leads to successful movies as they naturally attract a wider fan base. However, without these types of collaboration, the possibility for the movie being successful is still the case as we saw in Hugh Jackman and Ryan Reynolds. This indicates the correct role and correct cast can result in successful movies.

4.3 Correlation Analysis between Budget and Success Metrics

To examine the relationship between a movie's budget and its success, we conducted a correlation analysis. We calculated correlation coefficients between the budget and various success metrics, including box office performance, critical ratings, starring actors, movie duration, and release dates. This analysis allows us to determine the strength and direction of the relationship between budget and these success indicators.

Our findings reveal that while there is a positive correlation between budget and box office performance, the correlation coefficients are not particularly strong. This suggests that while budget may play a role in a movie's financial success, other factors also contribute significantly to the overall performance.

Additionally, we found that certain starring actors have a strong positive correlation with box office performance. Movies featuring popular and well-received actors tend to attract larger audiences and achieve higher box office earnings. Furthermore, movie duration and release dates also demonstrate some correlation with success metrics, indicating the importance of factors like runtime and timing in influencing audience reception.

4.4 Discussion of Findings and Implications

Based on the results and outputs, it is evident that budget is an important but not the sole determinant of a movie's success within the Marvel movie universe. While there is a positive correlation between budget and box office performance, the correlation coefficients indicate that other factors, such as the popularity of starring actors, movie duration, release dates, and quality of storytelling, play crucial roles in determining a movie's overall success.

These findings have significant implications for filmmakers and industry professionals. It highlights the importance of considering multiple variables and adopting a holistic approach when planning and producing Marvel movies. By focusing on factors beyond budget alone, such as selecting talented and well-received actors, crafting engaging narratives, and strategic release planning, filmmakers can maximize the potential for success within the Marvel movie universe.

Overall, this analysis provides valuable insights into the complex dynamics that contribute to the success of Marvel movies, emphasizing the need for a comprehensive understanding of the numerous variables that shape audience reception and financial performance.

5. Conclusion

5.1 Summary of Findings

In this study, we examined the factors influencing the success of Marvel movies beyond the budget. By analyzing a comprehensive dataset comprising information on budget, box office

performance, critical reception, starring actors, duration, and release dates, we gained valuable insights into the complex dynamics that shape the outcomes of these movies.

Our analysis revealed that while the budget is an important factor in a movie's financial success, it is not the sole determinant. Other variables, such as the popularity of starring actors, movie duration, release dates, and the quality of storytelling, also significantly contribute to a movie's success. We found that certain starring actors had a strong positive correlation with box office performance, and movies with compelling narratives and well-received performances could outperform higher-budget films in terms of critical reception and audience satisfaction.

5.2 Contributions and Implications

This study contributes to the existing literature by expanding our understanding of the factors influencing the success of Marvel movies. By considering a wide range of variables and conducting a comprehensive analysis, we demonstrated that budget alone does not guarantee a movie's success within the Marvel movie universe. This finding has implications for filmmakers and industry professionals, highlighting the need to adopt a holistic approach that considers multiple factors when planning and producing Marvel movies.

Our findings suggest that filmmakers should focus on selecting talented and well-received actors, crafting engaging narratives, and strategically planning release dates to maximize the potential for success. By considering these additional factors alongside budget, filmmakers can enhance the overall quality and reception of their movies.

5.3 Recommendations for Future Research

While this study sheds light on the factors beyond budget that influence the success of Marvel movies, there are several avenues for future research that can further deepen our understanding. Some potential directions for future research include:

Exploring the role of marketing strategies: Investigating the impact of marketing efforts, such as promotional campaigns, social media presence, and targeted audience engagement, on the success of Marvel movies.

Examining the influence of directorial styles: Analyzing the directorial styles and techniques employed in Marvel movies and their impact on critical reception and audience satisfaction. Investigating the effects of genre diversity: Exploring the influence of genre diversity within the Marvel movie universe and its impact on box office performance and audience engagement.

Considering the global market: Conduct a comparative analysis of the success factors for Marvel movies in different international markets, considering cultural variations and audience preferences.

By addressing these research gaps, future studies can provide a more comprehensive understanding of the multi-faceted factors that contribute to the success of Marvel movies, and by extension, inform decision-making processes within the film industry.

In conclusion, this study demonstrates that budget is not the sole determinant of success for Marvel movies. By considering a range of variables and adopting a holistic approach, filmmakers can maximize the potential for success and create engaging and well-received movies within the Marvel movie universe.

6. References

Lash, M. T., & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, *33*(3), 874-903.

Krauss, J., Nann, S., Simon, D., Fischbach, K., & Gloor, P. (2008). Predicting movie success and academy awards through sentiment and social network analysis.

Albert, S. (1998). Movie stars and the distribution of financially successful films in the motion picture industry. *Journal of Cultural Economics*, *22*, 249-270.

Bagella, M., & Becchetti, L. (1999). The determinants of motion picture box office performance: Evidence from movies produced in Italy. *Journal of Cultural Economics*, *23*, 237-256.

Tomaric, J. J. (2008). *The power filmmaking kit: make your professional movie on a next-to-nothing budget*. Taylor & Francis.

List of films based on Marvel Comics publications - Wikipedia. (2018, November 6). List of Films Based on Marvel Comics Publications - Wikipedia. https://en.wikipedia.org/wiki/List_of_films_based_on_Marvel_Comics_publications