Assignment 2 Report

Which Attributes Serve as Predictors of A Movie's Success on Netflix?

Author Information

Vincent Nguyen
1168263
COMP20008
qnng@student.unimelb.edu.au

Koquiun Li Lin
1319881
COMP20008
klilin@student.unimelb.edu.au

Isaac Tse
834150
COMP20008
itse@student.unimelb.edu.au

Christopher Kenneth Yapharis
1360684
COMP20008
cyapharis@student.unimelb.edu.au

Contents

Executive Summary

In this project, we analyze factors of a movie that affect the “success” of a movie, defined by its score on IMDb. The data was preprocessed with linear regression imputation, median imputation, and cleaning natural language with tokenization and stemming with Porter Stemmer. The data is then analyzed with a combination of box plots, bar charts and k-means clustering to observe trends or effects between variables and imdb_score.

These analysis has found 3 facts: actors and directors’ popularity is positively correlated with imdb_score; runtime does have an effect on imdb_score, but it is varied between different genre (in other words, there are interactions between runtime and genre); countries of production does not serve as a predictor, but instead there are many confounding effects.

However, this project has some limitations: only interactions between runtime and genre was explored, and only the top 5 most popular genres were considered in the analysis. Furthermore, only countries with more than 5 movies were considered in the analysis.

The project could be improved by looking at more interaction terms between genres and other factors. Having more data points would also potentially give more insight into the data.

Introduction

This report embarks on a comprehensive exploration into the factors that play a crucial role in predicting the success of movies on the Netflix platform. The research question guiding this analysis is: Which attributes serve as predictors of a movie's success on Netflix?

To conduct this investigation, we have leveraged data from the “titles” and “credits” CSV files. These sources provide information about movies, including their genres, production countries, and personnel involved in the filmmaking process. By examining this data, we aim to shed light on the multifaceted factors contributing to a movie's success on Netflix, ultimately providing valuable insights for both the film industry and the Netflix audience.

Methodology

Data Preprocessing and Wrangling

Firstly, actor/director names and IDs from credits.csv file were merged into the main titles.csv file using the movies’ IDs. As a result, it was found out that there are many missing actors/directors cells, but as these types of data cannot be easily imputed, cells with missing values are ignored for relevant analysis.

The missing data for imdb_score, imdb_votes, tmdb_score, tmdb_popularity is imputed using 2 simple linear regressions; a linear regression for tmdb_score with imdb_score as a (positive) correlation between the 2 types of score should be expected, and a linear regression for tmdb_popularity with imdb_votes, as the popularity metric should indicate that a movie is more talked about, hence somehow correlated with votes on IMDb.

The description of each movie is processed using natural language processing techniques: case stripping, removing punctuation appropriately, and stemming using Porter Stemmer. The paragraph is then tokenised and saved as a column in the dataframe.

Considering the research question, we decided to focus on specific predictors that might provide an effect towards the success of a movie. These include runtime, genres, countries of production, actors and directors. We posit that the imdb_score serves as a significant indicator of a movie's success, given its reliance on a well-established database with authoritative reviews, and comprehensive cast and crew information (Lewis, 2023).

Runtime in Correlation With Genre

Runtime was compared in conjunction with genres, as Figure 1 shows every genre has a different average runtime. To provide a sample that ensures no exaggerated number of figures, only the top 4 genres that have the highest average imdb_votes will be focused on (Figure 2).

A high average imdb_votes indicates that its imdb_score will be more accurate. Due to this fact, the relationship between runtime and imdb_scores across all top 4 genres will be analyzed using Pearson correlation.

Furthermore, to verify the results of Pearson correlation, k-means clustering was also employed. VAT is used to find the number of clusters. Through clustering, how each top 4 genres categorizes the movie's runtime can be qualitatively quantified. The centroid of genres' clusters are calculated to determine which runtime cluster provides the highest imdb_score.

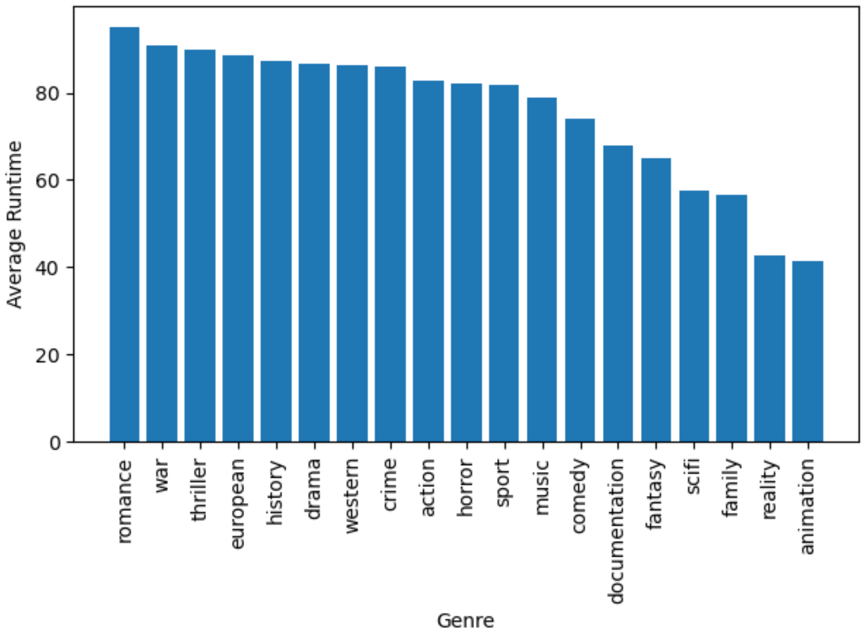


Figure 1: Bar plot of average runtime of all genres

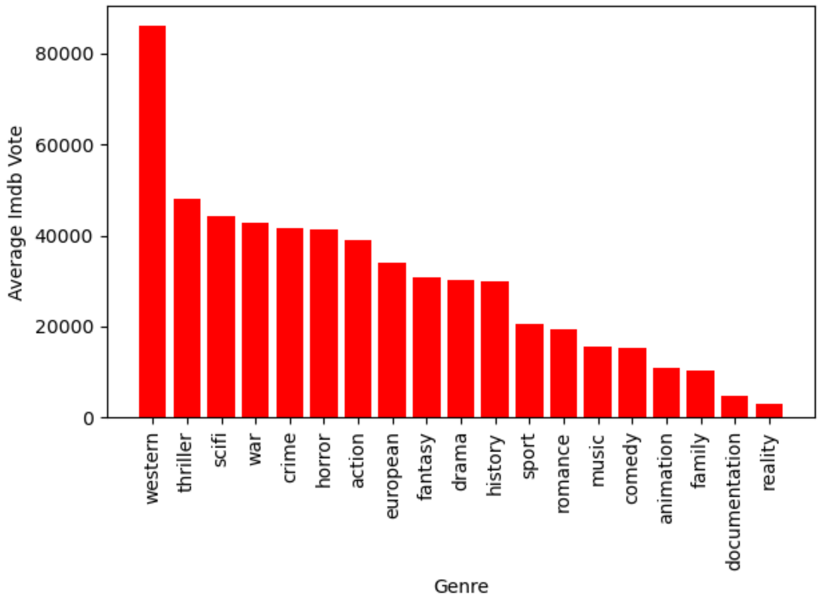


Figure 2: Bar plot of average imdb_votes of all genres

TF-IDF

As a tangent, an analysis on the descriptions for the movies was performed. With a cleaned description and tokens, TF-IDF was employed to obtain important keywords for the top 5 most popular genres.

For each considered genre, all of the movies tagged with the genre will have its description added to the corpus. Then, using TF-IDF vectoriser from the sklearn library, the parse matrix containing every token's TF-IDF scores are computed. The top 3 tokens with the highest TF-IDF metric are added to a dictionary, counting how many times said tokens have been considered "top 3". The top 5 most repeated tokens are then considered the "keywords" for the genre.

A drawback of this way of considering keywords is that tokens with tied frequency may not be considered as keywords due to being knocked out of the top 5 by arbitrary means.

Production Countries

Analysis approaches include calculating average imdb_score for all of the production countries, then sorting and plotting the top 10 groups of production countries with the highest average scores.

Further investigation involved examination on the top 10 individual production countries, ordered by their average imdb_scores without considering collaboration from other nations.

The dataset is later filtered to identify countries with substantial movie production, examining their mean imdb_scores through box plots on the same axis, with outliers removed (as only the general trend of imdb_score is of interest) and ordered by their median.

Actors and Directors

The Pearson correlation coefficient between the imdb_score of a movie and the average “actor scores” of its cast was used to quantify the effect of actors/directors experience and ability on the success of movies/TV shows.

“Actor score” was computed by taking the average imdb_score of all movies a particular actor appeared in, and similarly for “director score”.

Only actors with more than five appearances, and directors with more than two were considered in this analysis. This is because actors or directors with only a single appearance significantly skew the data, as in many of these cases this is due to a movie having actors who only appeared in that one movie, which leads to a meaningless perfect correlation.

It should be noted, however, that doing this greatly reduced the sample size of actors and directors in the data.

Data Exploration and Analysis

Linear Regression

A scatter plot is first used to determine at a glance if any transformation to the data is necessary:

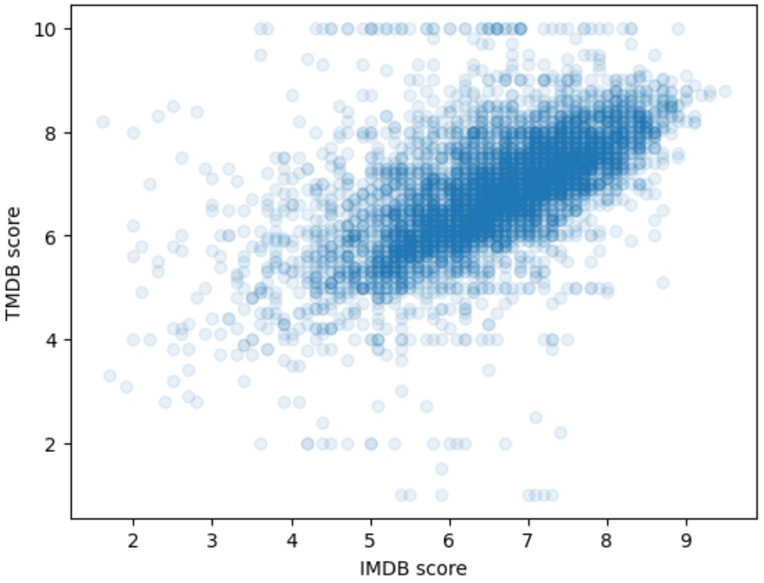


Figure 3: Scatter plot of tmdb_score vs. imdb_score

It can be seen that most data points are located along a linear line. From this, no transformation is necessary. All of the known pairs (imdb_score, tmdb_score) are extracted, then split into 80% training set and 20% testing set. Using the training set, a linear regression is fitted.

The MSE for the training set is 0.897, with the MSE for the testing set being 0.836. The low MSE suggests that the model has good accuracy, along with training MSE and testing MSE being very close to each other suggests that the model is consistent in its prediction, thus is ready to be deployed.

Missing data is then imputed using this model for cases where 1 value is known and the other is not. After this, for cases where 2 values are unknown, median imputation will be performed.

Similarly, before fitting a linear model for imdb_votes and tmdb_popularity, a scatter plot is used to explore the data first:

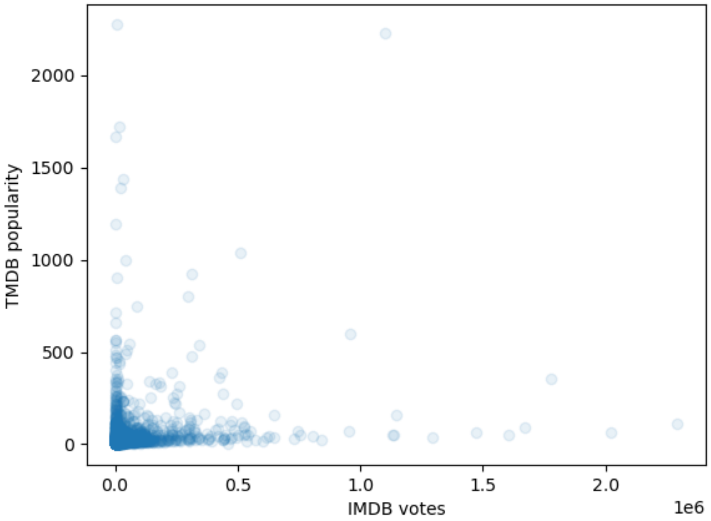


Figure 4: Scatter plot of tmdb_popularity vs. imdb_votes

It is observed that both the popularity and the number of votes are positively skewed. To observe these data better, a log (base 10) transformation was applied to both sets of data:

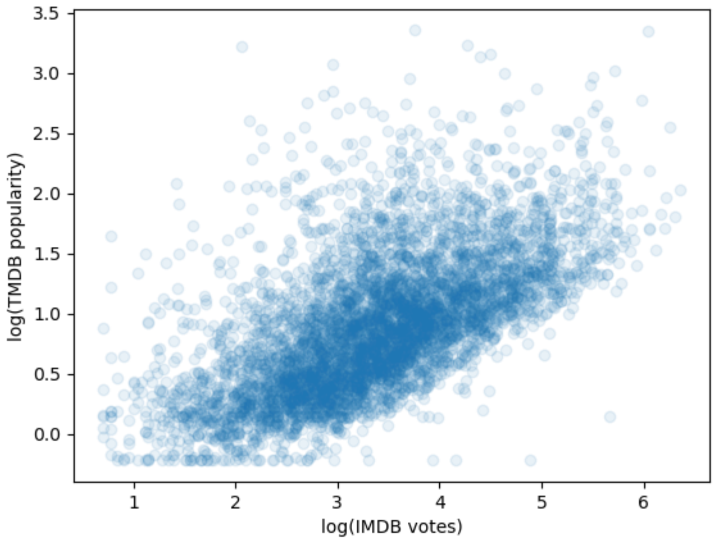


Figure 5: Scatter plot of log(tmdb_popularity) vs. log(imdb_votes)

There is a clearer relationship between these 2 transformed data. Similarly, the data are split into training and testing sets, with a ratio of 80:20.

The MSE for the training set is 0.195, with the MSE for the testing set being 0.205. Again, this suggests that the linear model is consistent, and can be safely deployed to cases where one value is known and the other is unknown. As the numbers obtained from the data are transformed, during the imputation, the imputed value will be transformed back by taking 10^(value). After this, cases where both values are missing will be imputed by the median of the column.

Runtime in Correlation With Genre

As shown by Figure 2, the top 4 movie genres are western, thriller, scifi, and war. To get a preliminary understanding of how runtime is correlated to high imdb_score for the top 4 genres, we created 4 scatter plots and calculated the respective Pearson correlation. These scatter plots will also show the clustering of movie length that will be attended in the later analysis.

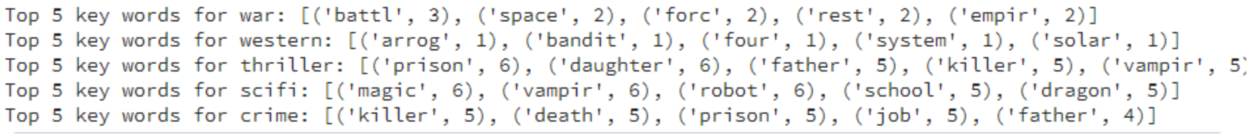
As shown by Table 1, The only genre that shows an insignificant 0.00326 correlation is the western genre. On the other hand, thriller, scifi, and war have low negative linear correlations. This means that for these 3 genres, runtime is more influential on the imdb_score.

Table 1: Pearson correlation of runtime vs imdb_score

	Western	**Thriller**	**Scifi**	**War**
Correlation	0.00326	-0.21036	-0.249017	-0.145516

Keywords for the top 5 most popular genre

While this does not directly answer the research question, we found the following through applying TF-IDF techniques



```
Top 5 key words for war: [('battl', 3), ('space', 2), ('forc', 2), ('rest', 2), ('empir', 2)]
Top 5 key words for western: [('arrog', 1), ('bandit', 1), ('four', 1), ('system', 1), ('solar', 1)]
Top 5 key words for thriller: [('prison', 6), ('daughter', 6), ('father', 5), ('killer', 5), ('vampir', 5
Top 5 key words for scifi: [('magic', 6), ('vampir', 6), ('robot', 6), ('school', 5), ('dragon', 5)]
Top 5 key words for crime: [('killer', 5), ('death', 5), ('prison', 5), ('job', 5), ('father', 4)]
```

Figure 6: The top 5 keywords for the top 5 most popular genres. The number associated with each word signifies the frequency the keyword was deemed important in a description.

The data obtained for western movies are not very useful; all keywords are only considered important once. This is due to the small sample size of western movies.

For other genres, many of the tokens are associated with the genre semantically, as expected.

Production Countries

According to Figure 7, the graph shows the top 10 production countries by average imdb_score. Generally, most of the top 10 production countries are relatively wealthy countries. This suggests that there may be a correlation between wealth and the production of high-quality films. The United States ('US') and The United Kingdom ('GB') are clear leaders in collaboration with various countries, but there are a number of other countries that produce high quality movies. Surprisingly, there is some evidence to suggest that countries that produce movies together tend to produce movies together to produce successful movies. Hence, co-producing movies with other countries can be a good way to produce successful movies, but it is not a guarantee of success.

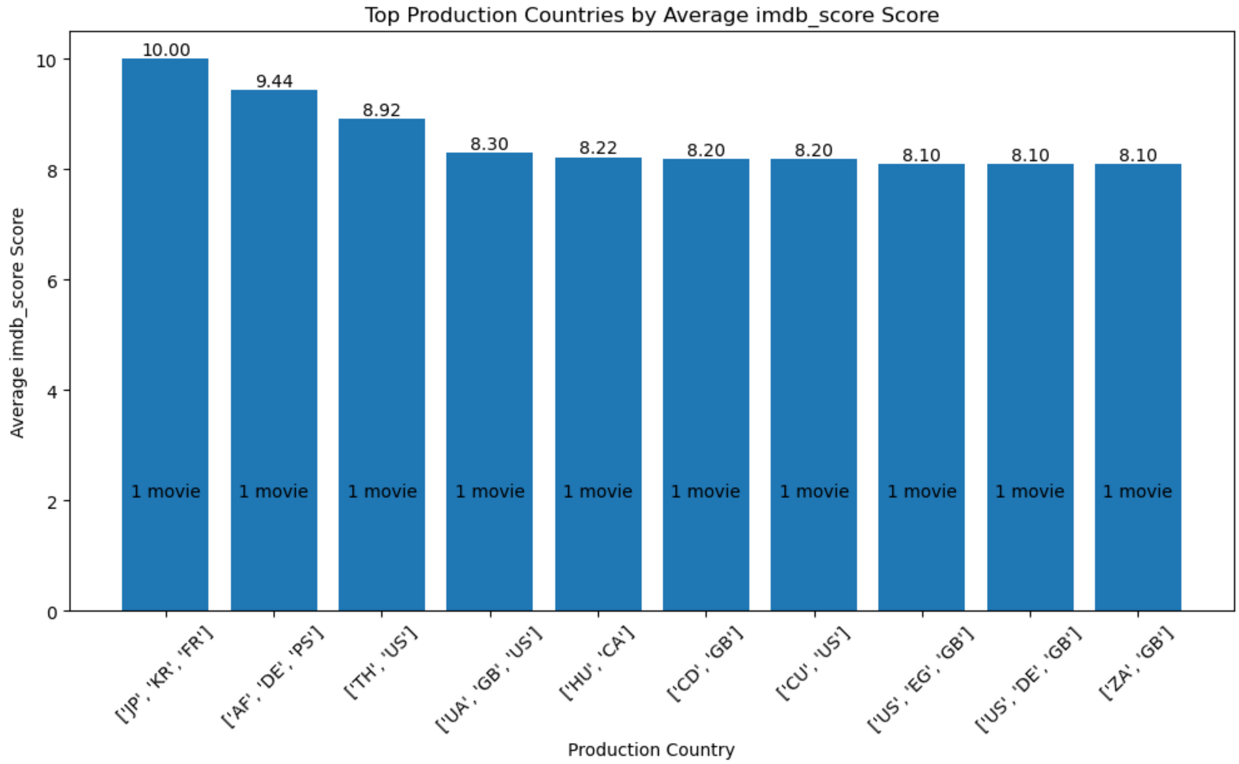


Figure 7

Examining the top 10 individual production countries, ordered by their average imdb_score without considering collaboration from other nations. While wealthier countries frequently participate in international projects to make successful movies, interestingly, from Figure 8, mostly these individual and niche countries, as evidenced by their high imdb_score, have an advantage in producing successful movies independently. Afghanistan ('AF'), Puerto Rico ('PR'), and Portugal ('PT'), which are not typically considered major players in international collaborations, have achieved high average imdb_score (7.7), despite producing a limited number of movies (1-2). It could be because the viewer base is smaller, and the niche movies tend to align more with their target audience. This conclusion highlights the diversity in the movie industry and indicates that success is not solely determined by the economic power of a country. Furthermore, the observation that countries with higher production volumes, such as Israel ('IL') and China ('CN'), tend to have lower average imdb_score is noteworthy. This observation hints at a potential relationship between the scale of production and movie quality. It is possible that countries with extensive film industries may produce a wider spectrum of films, including some of lower quality, which can influence their average imdb_score.

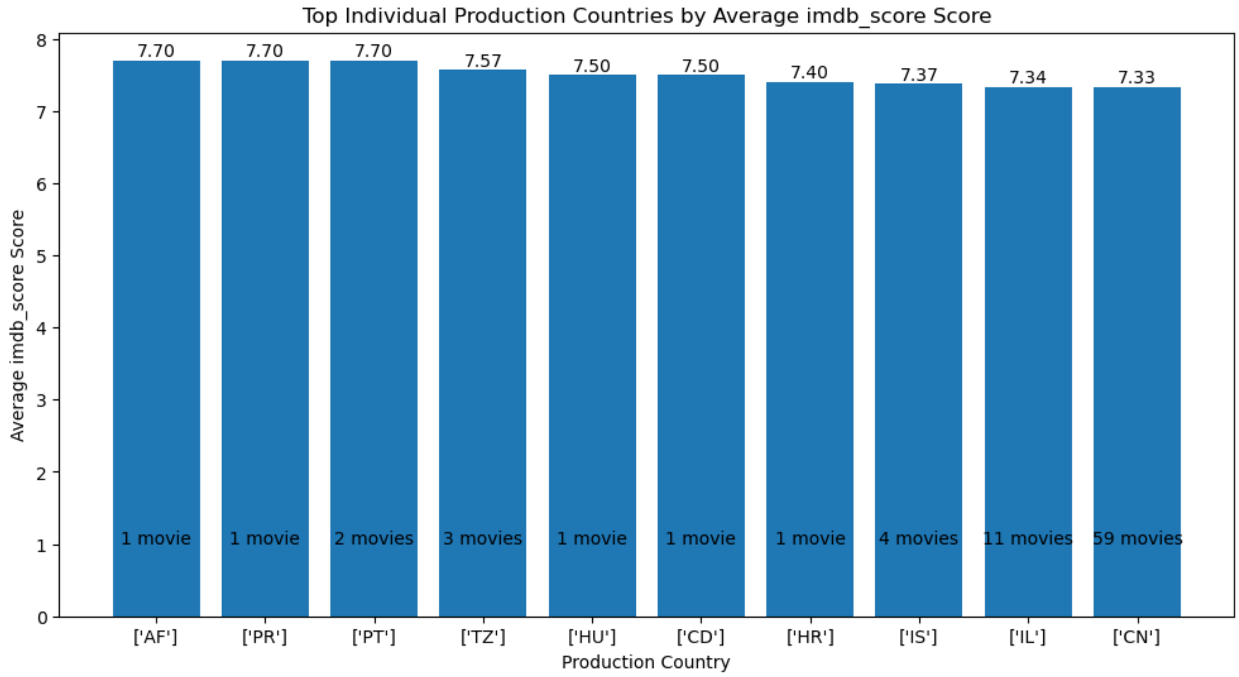


Figure 8

Since countries with low movie counts have a lot of variance for their mean, we decided to focus on countries which have a considerable number of movies produced. A parallel box plot focusing on the imdb_score for countries producing over 5 movies was generated. Figure 9 shows the distribution of the number of movies produced by various individual countries. The figure illustrates the overall median is 6.57 movies produced per country. Countries such as Korea (‘KR’), China (‘CN’), the United States (‘US’) and India (‘IN’) have relatively large spread, potentially a reflection of the large amount of movies produced, inevitably widening the spectrum of movie quality.

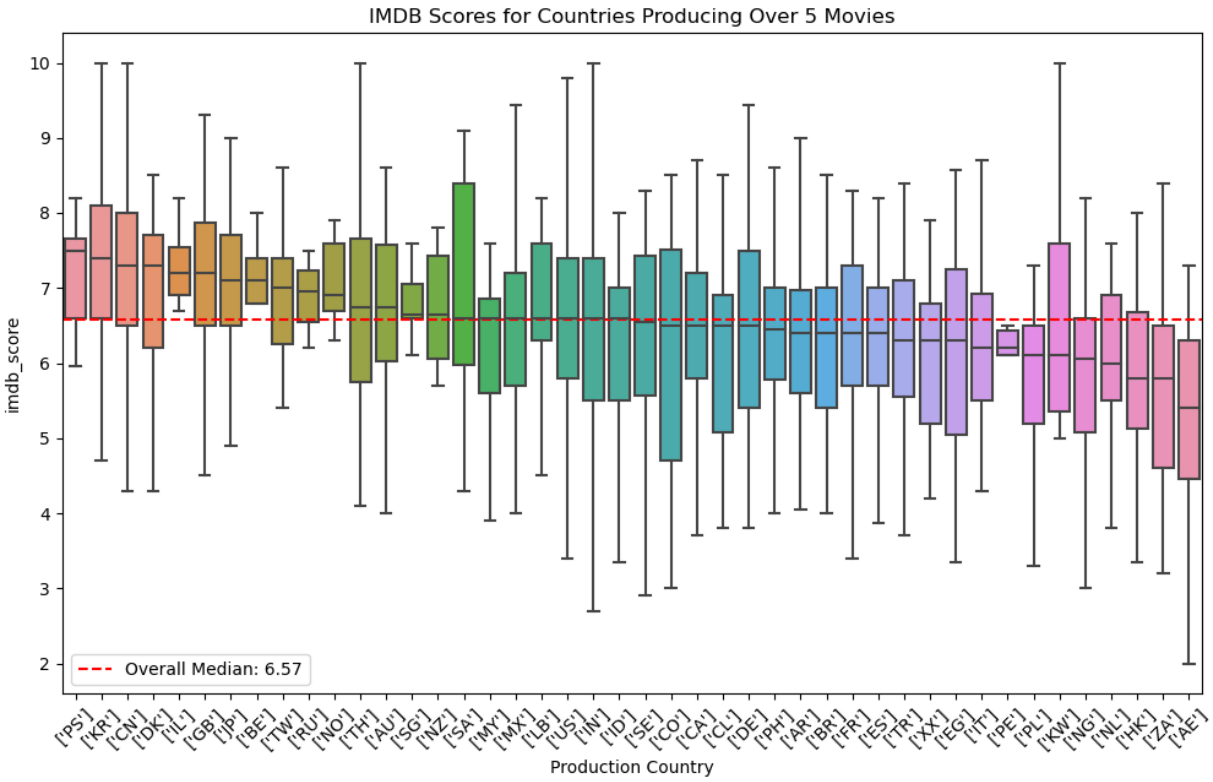


Figure 9

Actor and Director

For actors, the Pearson coefficient turned out to be 0.44, a moderate correlation. This means there is some relationship between better actors leading to better movies, but the effect is likely muted by the large number of actors in each movie. The correlation may also be complicated by a lack of data related to actors - there is no way of knowing the 'expected' impact of each particular actor, as for example it would be expected that the main cast would have a far greater effect than any background actors.

For directors, the coefficient was 0.93, which implies a very strong correlation. This means it is very likely that having a director who has previously directed multiple high scoring movies will have some positive effect on a movie.

Table 2: Pearson correlation of imdb_score and average actor scores

	Actors	**Directors**
Correlation	0.44	0.93

Results

For runtime in correlation with genre, a further analysis using clustering is done. Again, the same pattern as the pearson correlation analysis occurs. Western genre uniquely has 4 clusters. In general, the clusters are short, medium, and long runtime. However, there is a further sub-clustering between high-score and low-score medium movies, due to the clusters in the middle being on top of one another.

From the calculation of centroid of clusters, both of the sub-clusterings of medium runtime have a relatively low centroid for imdb_score when compared to clusters of short and long runtime. This could mean that there is a negative effect on classifying the movie as medium runtime.

For the other 3 genres, the movies can be classified into 2 clusters which are short and long runtime. This pattern could mean that the majority of directors, disregarding the genre of the movie, do not want their movie to be in the awkward position of medium runtime. Medium runtime could lead to the inability to tell the full story or drag out the story, making it less enjoyable and possibly lowering the imdb_scores. This behavior is shown in Figure 10 and Table 3 of western genres that have a medium runtime cluster.

From the calculation of centroid of clusters, it is evident across these 3 genres that the clusters of short runtime always have a higher centroid of imdb_score, showing the relationship that the shorter the runtime, the higher the imdb_score of the movies.

Table 3: Pearson correlation of runtime vs imdb_score

	Western	**Thriller**	**Scifi**	**War**
Correlation	0.00326	-0.21036	-0.249017	-0.145516

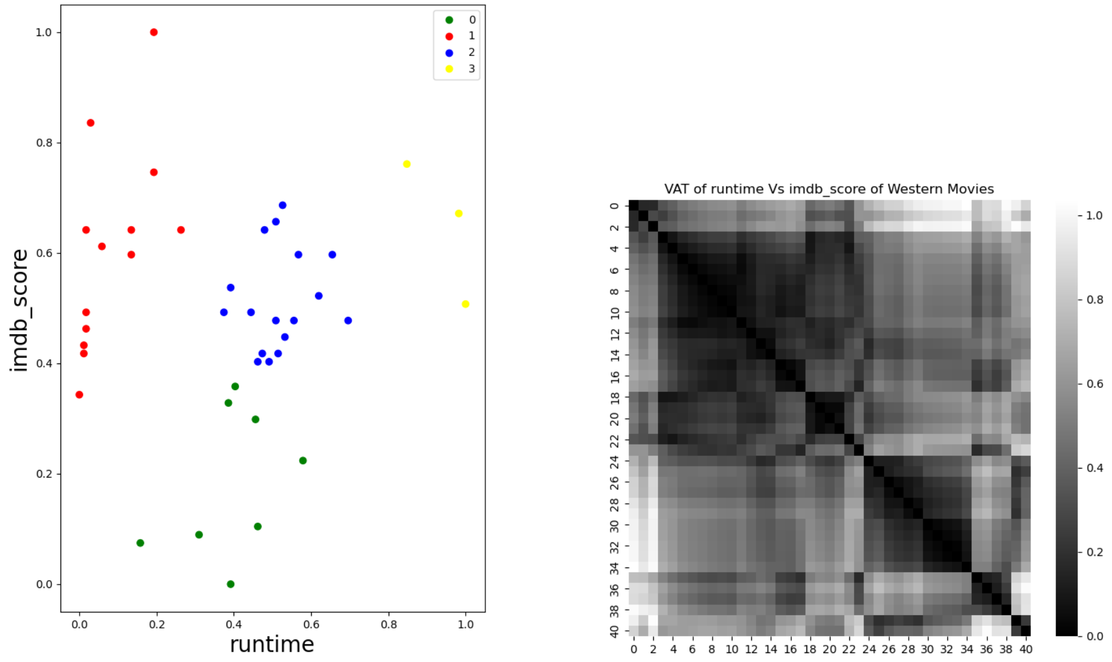


Figure 10: Clustering of runtime vs imdb_score of western movies with k = 4

Table 4: Centroid of western genre

Cluster	runtime	imdb_score
0	0.393275	0.184701
1	0.083221	0.605052
2	0.517716	0.514486
3	0.943470	0.646766

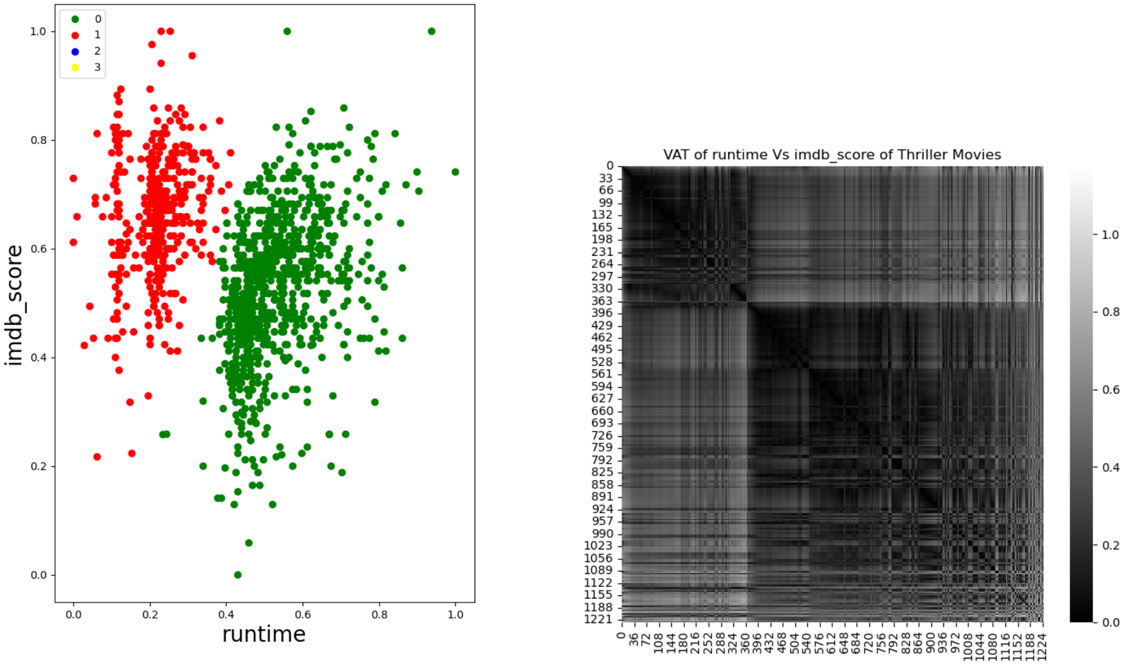


Figure 11: Clustering of runtime vs imdb_score of thriller movies with k = 2

Table 5: Centroid of thriller genre

Cluster	runtime	imdb_score
0	0.536454	0.526848
1	0.216821	0.662874

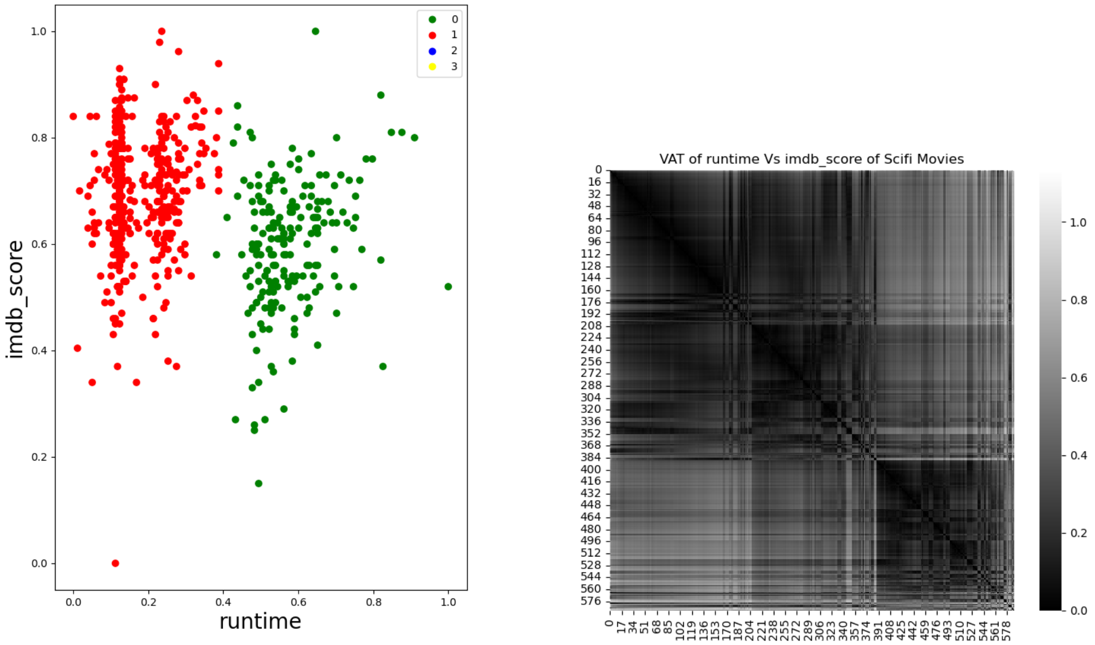


Figure 12: Clustering of runtime vs imdb_score of scifi movies with k = 2

Table 6: Centroid of scifi genre

Cluster	runtime	imdb_score
0	0.579821	0.596120
1	0.178901	0.692004

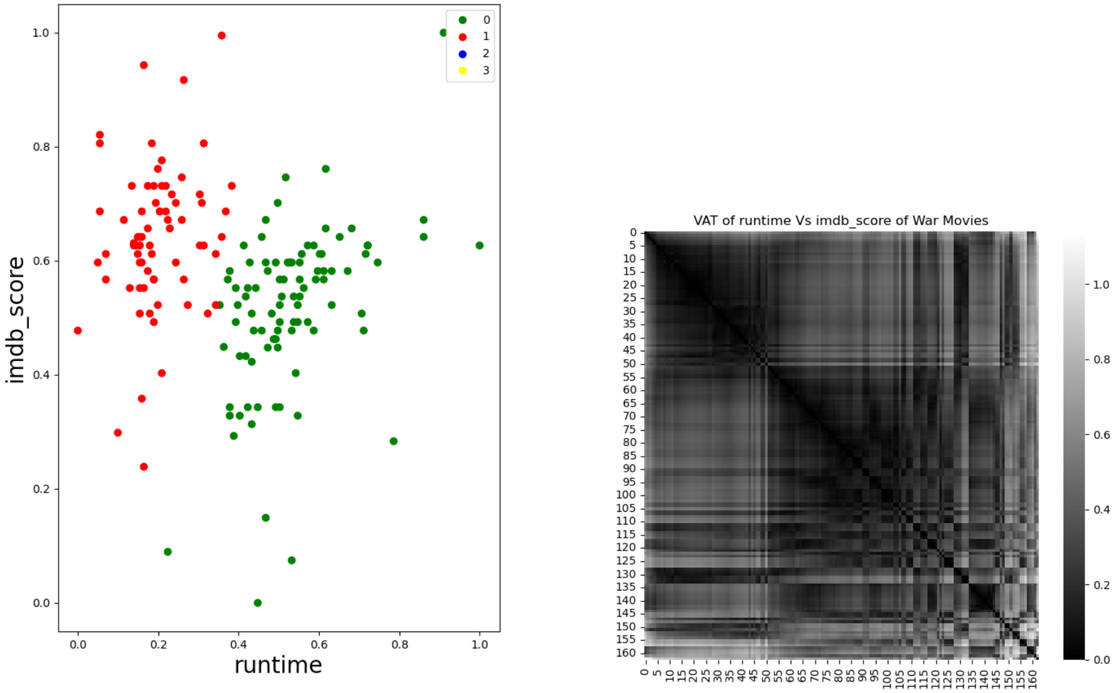


Figure 13: Clustering of runtime vs imdb_score of war movies with k = 2

Table 7: Centroid of war genre

Cluster	runtime	imdb_score
0	0.533041	0.514769
1	0.197674	0.639600

One complication encountered is when calculating Calinski-Harabasz Indexes, the best clustering number for western and thriller genres are different then what are used in this analysis. This shows the effect of different interpretation of the VAT graphs.

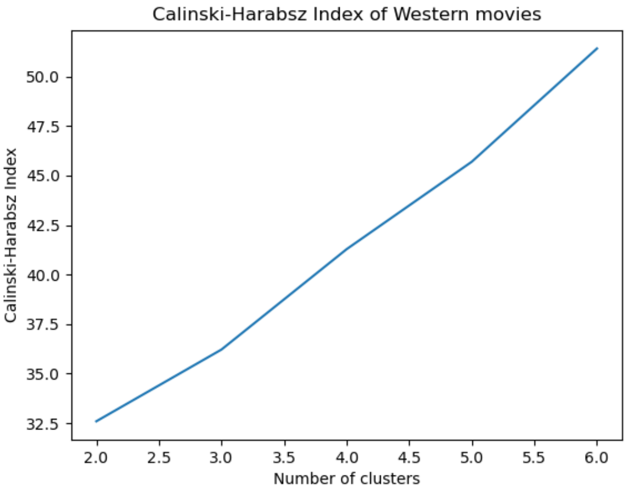


Figure 14

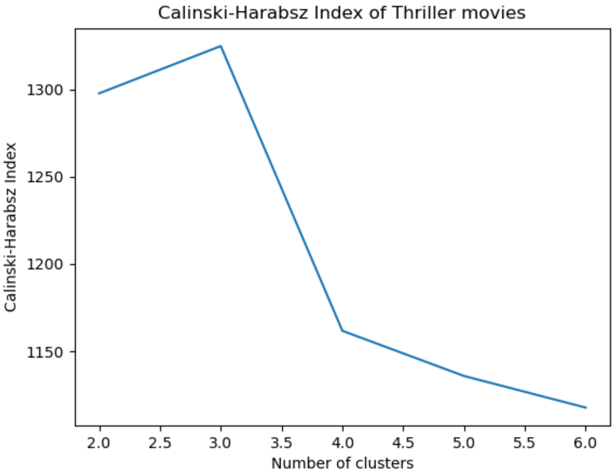


Figure 15

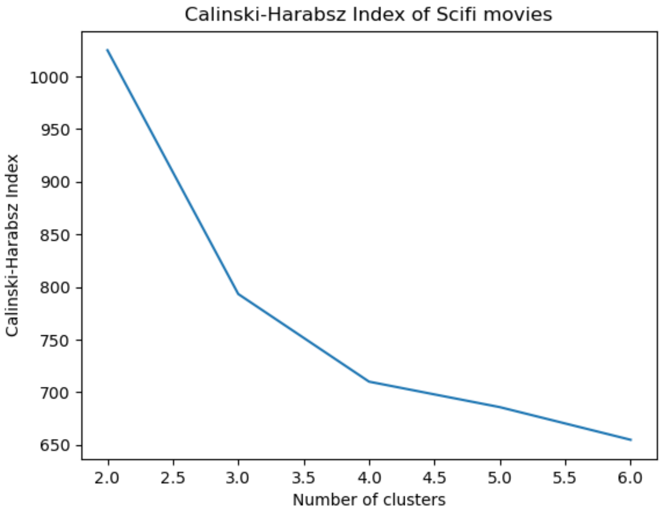


Figure 16

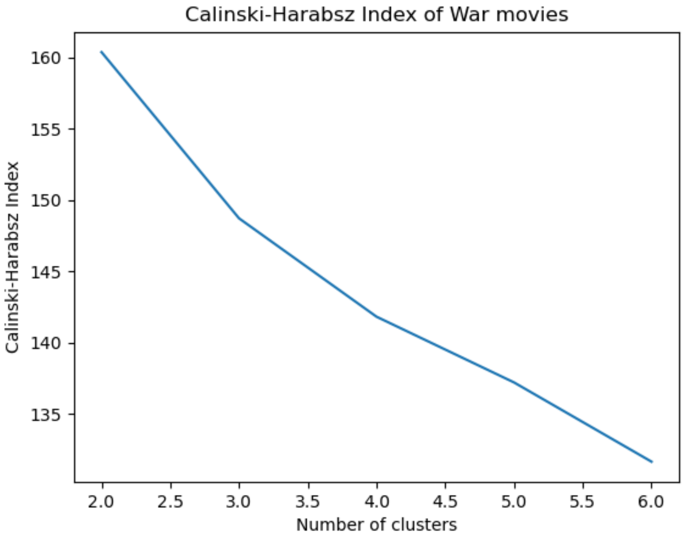


Figure 17

Findings and Interpretation

The findings from our analysis reveal several key insights into the factors influencing the success of movies.

In terms of runtime,

- For 3 out of 4 selected genres, there is a low negative linear correlation between runtime and imdb_scores. This means that as runtime becomes longer, there are some decreases in the imdb_score of the movies.
- Coincidentally, the same 3 out of 4 genres are each divided into clusters of short and long runtime, where short runtime has a higher centroid of imdb_score compared to long runtime. These show the reluctancy of having an awkward medium runtime and the tendencies that short movies get higher scores.

Due to the small correlation between runtime and imdb_score, movie runtime is not a definitive predictor of imdb_score across various genres, However, this predictor should not be dismissed entirely. The consistency in results across different analytical methods suggest potential use in certain contexts.

In terms of production of country,

- Rich countries benefit from international collaboration, while niche countries can thrive independently.
- The size of a country's film industry also matters. Extensive industries produce a broader spectrum of films, impacting a movie's success.

To succeed, countries should focus on international partnerships, tailor productions to specific audiences, and prioritize quality over quantity. This challenges the idea that a country's production capacity alone predicts a movie's success, offering insights for movie industry professionals and researchers. Unexpectedly, success is a multifaceted interplay of factors that cannot be summarized by production of countries alone.

In terms of actors and directors,

- There is a strong correlation between a movie’s imdb_score and the average success of its directors, which strongly implies that getting a successful director is a vital part of making a successful movie/tv show
- There is a moderate correlation between a movie’s imdb_score and the average success of its actors, implying that while it is important to have good actors, the effect is diluted compared to directors, likely due to the larger number of actors per show.

Limitations and improvement opportunities

There are some limitations that we encountered during this analysis. Firstly, the results of these projects are limited due to only considering specific predictors. Some other predictors might be more effective in determining high imdb_score, which can be analyzed further by future studies. Secondly, the different interpretations of the number of clusters shown by the VAT might cause confusion and inconsistent representation of the result of the analysis. Future studies can solve this issue by focusing more on Calinski-Harabasz Index. Furthermore, the underpinning assumption is regarding IMDb metrics being the better fit compared to TMDB metrics on measuring movie success, which was decided somewhat arbitrarily using external information that is not verified by the provided data.

Conclusion

In summary, our analysis not only sheds light on the key factors influencing movie success on Netflix but also emphasizes the multifaceted nature of this interplay. These insights have the potential to reshape strategies in the movie industry and guide future research endeavors, demonstrating the value of our analysis in understanding the dynamics of success in this context. Given the nuanced nature of these findings, we recommend that filming industry professionals prioritize the recruitment of exceptional directors and prioritize the quality of movies over quantity. This focus on great directors and a commitment to producing high-quality content can significantly enhance the chances of success, even in a diverse and competitive landscape. While our analysis serves as a foundational step, future studies should explore more extensive socioeconomic factors and intricate interactions between variables to further refine strategies in the ever-evolving world of streaming entertainment.

References

Lewis, R. (2023, October 3). IMDb. Encyclopedia Britannica. https://www.britannica.com/topic/IMDb