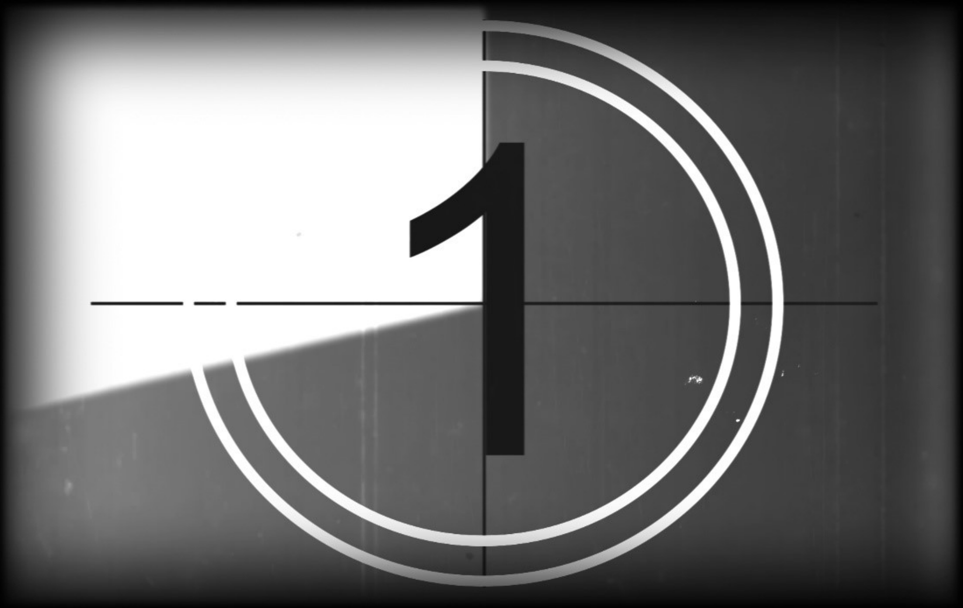


COSC3000 S1 2023

VISUALISATION REPORT:

IMDB DATA

ERIC SHEN
46466369

Contents

Introduction

> *Every viewer is going to get a different thing. That's the thing about painting, photography, cinema.*
>
> – David Lynch, Director of Twin Peaks

From the Lumière brothers' legendary public screening of their short films in Paris to modern classics such as The Shawshank Redemption and Pulp Fiction, cinema has captivated the public eye for over a century since its inception. Filmmaking industry giants the likes of Disney and Warner Bros have constantly vied for the audience's favour with every new movie released. This has led film analysts to investigate various aspects of films in order to better predict and understand the trends behind the success of one film and the failure of another.

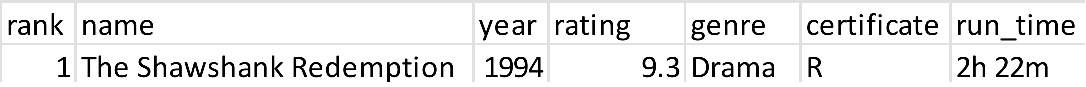
The Internet Movie Database (IMDb) is a comprehensive online public database storing information related to film and television. This report focuses on the top 250 movies of all time and aims to investigate the relationships between various aspects of each film to determine if there are common key factors behind them. Additionally, due to the data spanning several decades, variations over time will be of interest. The investigation will be primarily achieved using data visualisation methods to accentuate trends and patterns as an exploratory tool.

The Data

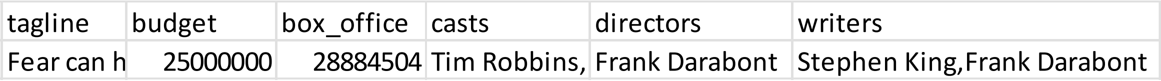
The following parameters were recorded for each movie:

- Rank
- Name
- Year
- Rating
- Genre
- Certificate (age rating)
- Runtime
- Tagline
- Budget
- Box office
- Cast
- Director(s)
- Writer(s)

For example, the following is the first entry of the dataset:



rank	name	year	rating	genre	certificate	run_time
1	The Shawshank Redemption	1994	9.3	Drama	R	2h 22m



tagline	budget	box_office	casts	directors	writers
Fear can h	25000000	28884504	Tim Robbins,	Frank Darabont	Stephen King,Frank Darabont

The data was downloaded in the form of a comma-separated values file from Kaggle (Raju C., 2023), but was originally collected by scraping the data from the IMDb website. Budget and box office are measured in USD, and the rating is a number from 1 to 10 with one decimal place. To standardise entries in the data for this project, any movies with inconsistent values or missing data for certain variables were dropped from the dataset. The certificates were standardised to their US equivalent and the run_time was formatted to be in the form #h ##m. The run time in minutes was later calculated as the column run_min. Movies with box office totals deemed unusual compared to their budgets were researched individually and rectified to their correct values manually. In most cases, the listed box office was magnitudes lower than the budget, indicating an anomaly given the popularity of the films in this dataset.

These adjustments reduced the total movie count from 250 down to 203.

Data Visualisation and Analysis

As there was a small amount of numerical variables, a pairplot was deemed to be a reasonable starting point as it allows for a broad overview of the data for any potential trends. The scale for budget and box_office axes were altered to be logarithmic, as otherwise outliers caused the plot to be difficult to interpret.

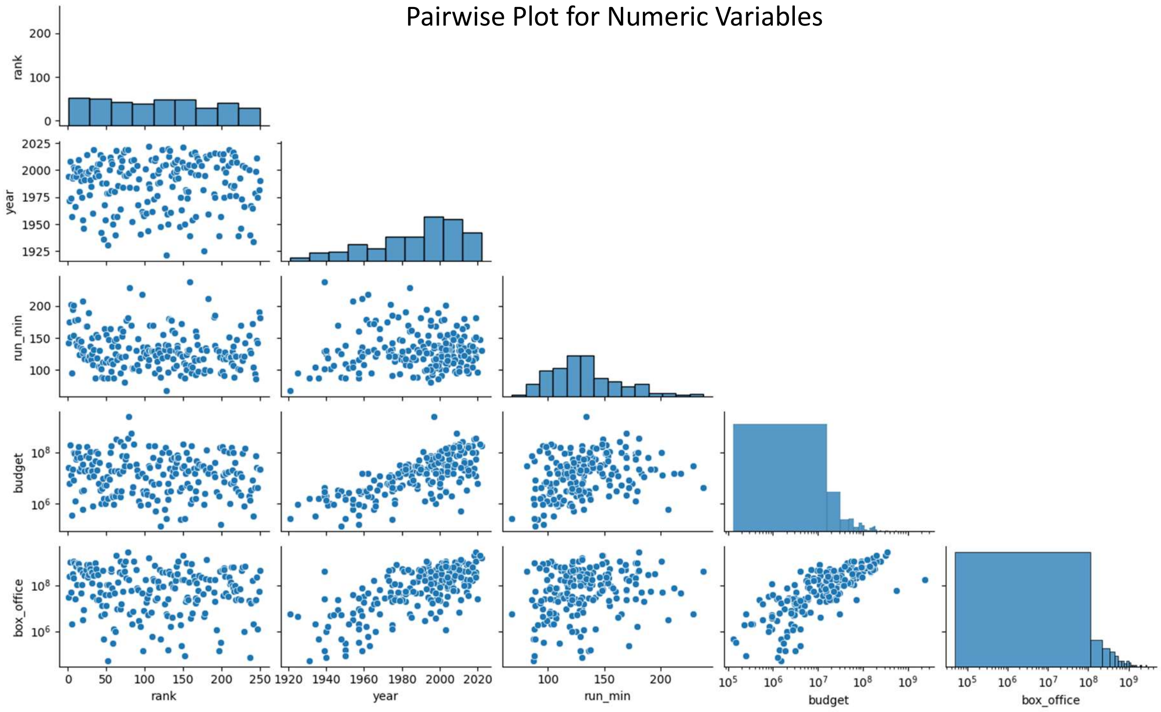


Figure 1 *Pairwise relationships between various numerical elements of IMDb film data. From top to bottom and left to right: rank, release year, run time (minutes), budget and box office sales*

From the leading diagonal in Figure 1, we can investigate the univariate distributions of the variables. This lets us see that there are slight skews in the distributions of release year and run time, as well as significant positive skew for both box office and budget. If the dataset had not been standardized prior to the visualisation, the rank would have been uniform, since there is exactly one film of each rank. The skew of release years can likely be attributed to the general increase in popularity of filmmaking and cinema as a whole.

When examining the bivariate plots, there appears to be a positive correlation between budget and box office; year and budget; and year and box office. Additionally, run time seems to have decreased over the years, becoming densely concentrated between 100-150 minutes. Interestingly, rank appears to have no effect on any of the other variables.

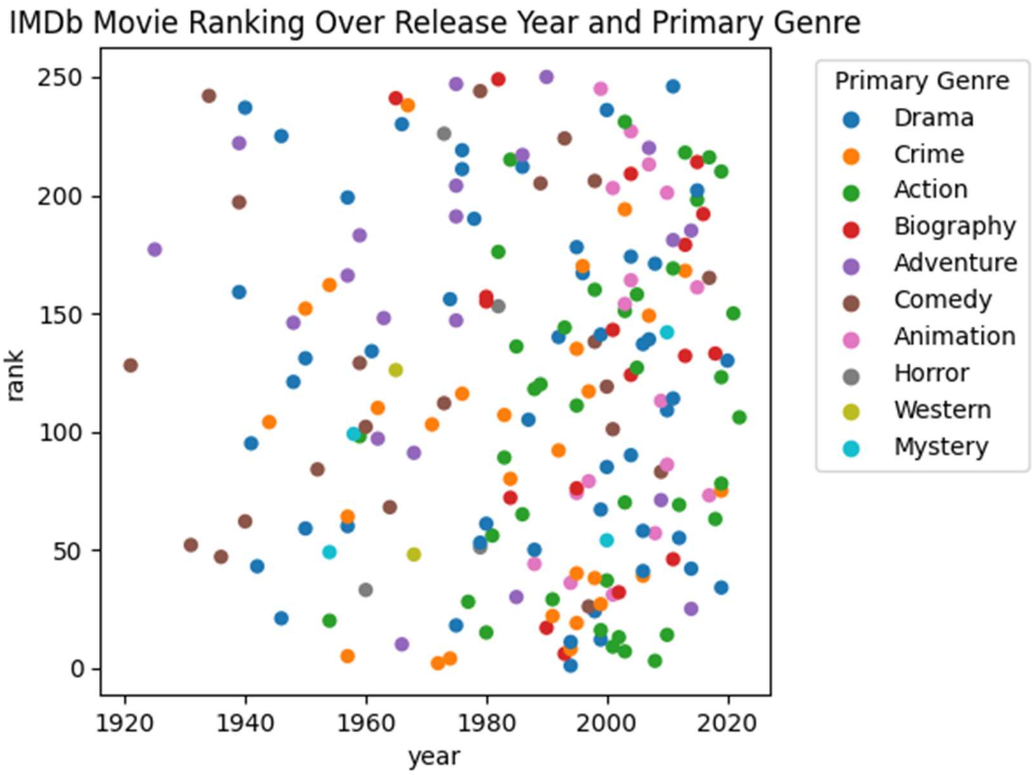


Figure 2 *Variation of genre density over time*

Figure 2 takes the neutral relationship between year and rank and investigates whether an underlying relationship exists between the two and a film's primary genre. While the large number of distinct genres causes this plot to be difficult to interpret, closer observation yields that a large number of action films post-1980 are in the top 250. This could be due to a lack of action movies prior, although evidence suggests that the onset of computer generated images (CGI) in the 1980s may have been the catalyst that allowed filmmakers to be more creative in their effects (Formichella, 2020). The same time period also saw the increase in popularity of animation, another genre that benefited from the advancements of CGI.

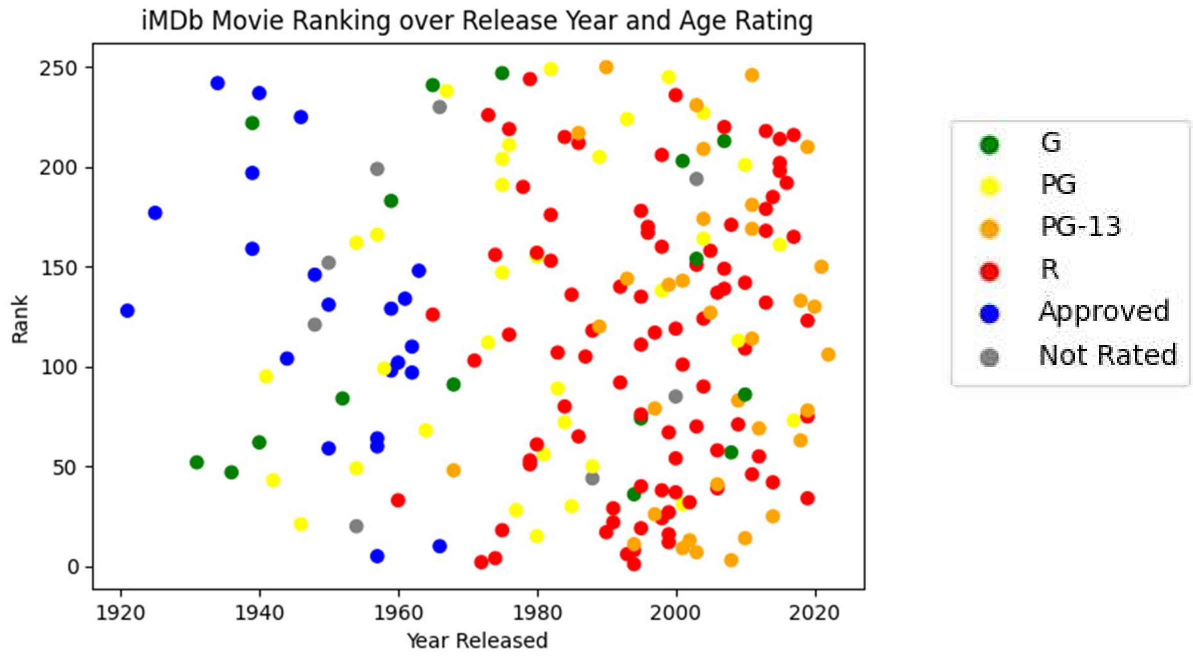


Figure 3 *Relationships between age ratings and release year*

This visualisation benefits from the reduced class count of certificates compared to genres, and demonstrates the shift in film certification over time, where R, PG-13 and PG films became much more prevalent. This is due to the Motion Picture Association of America (MPAA) establishing the current rating system in 1968, as part of a plan to provide parents with a guide for child appropriate film and television content (Dow, D. 2009). Figure 4 (below) is another alternate visualisation of this trend, a bar plot with 20-year bins. Here, the sudden increase in R rated films compared to before is obvious.

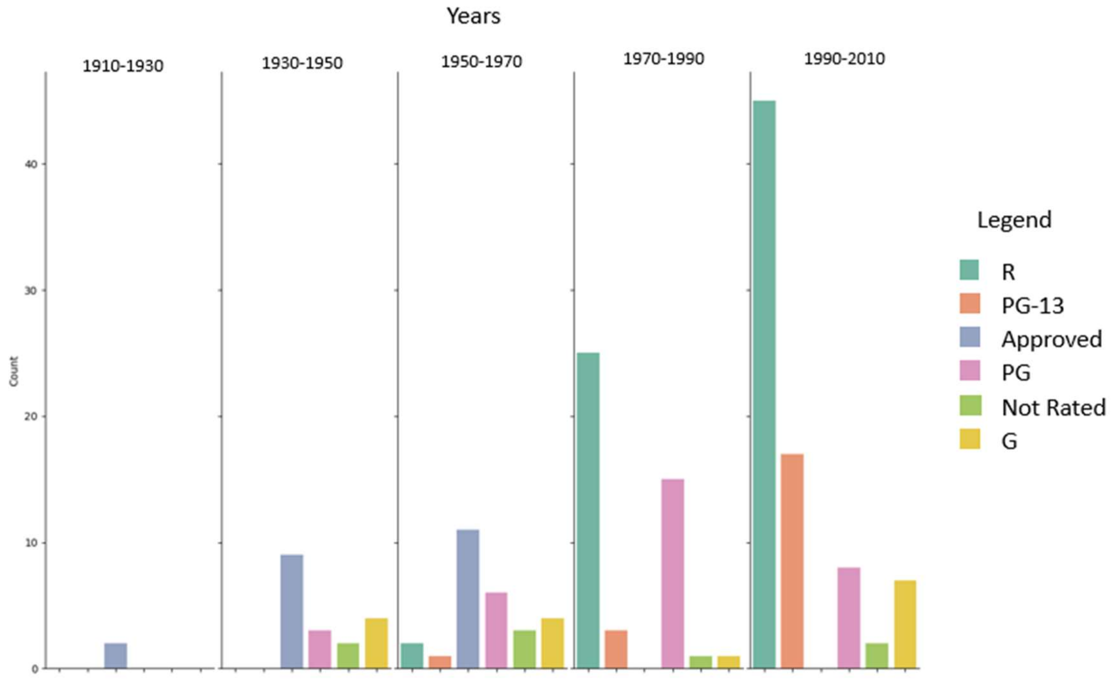


Figure 4 *Categorical bar plot of relationships between age ratings and release year*

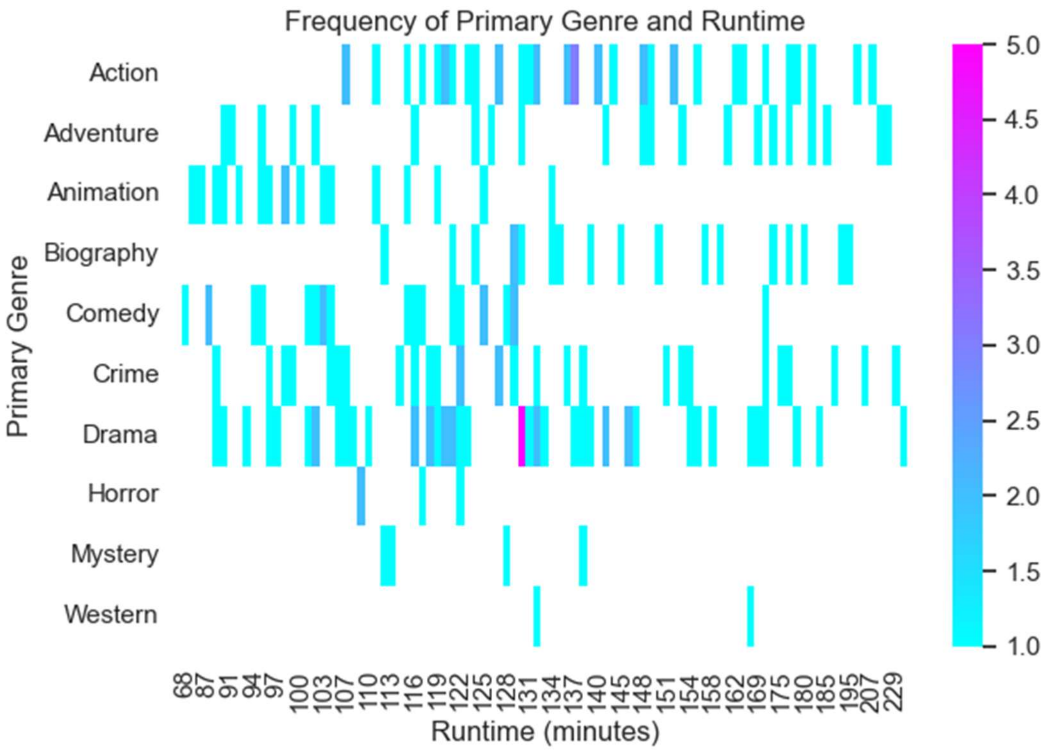


Figure 5 *Heatmap illustrating the frequency of genre-runtime pairs*

In order to investigate the possibility of a relationship between a genre and its run time, a heatmap was employed. This allows us to easily see groupings of datapoints, as present in the several of these genres, with animation favouring shorter run times as opposed to biographies which are centred closer to the longer run times. Interestingly, dramas, crime and adventure films span the widest range of run times. This implies that run time does not play an integral part in a film from those genres' success.

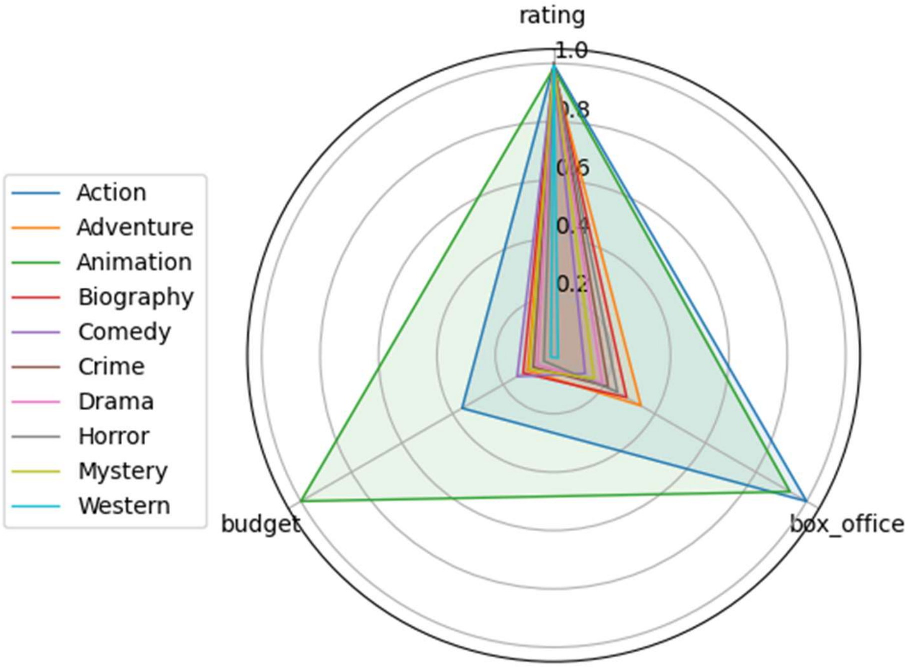


Figure 6 *Radar plot of each genre and its average rating, budget and box_office. The values were normalised within [0, 1] to aid the illustration.*

In an attempt to visualise the average rating, box office and budget of each genre, a radar plot was chosen as it could be useful in clearly showing the difference between each genre. Unfortunately, the average values of each genre were very similar to each other with the exception of the budget and box office of animation and action films. The normalisation method may have also played a part in making the graph convoluted, nearly resulting in every rating converging to one value. Despite these issues, the graph illustrates the high budget and box office of animation films, and the relatively high return studios gain on average from action films. In particular, animation seems to have the worst budget:box office ratio of the genres.

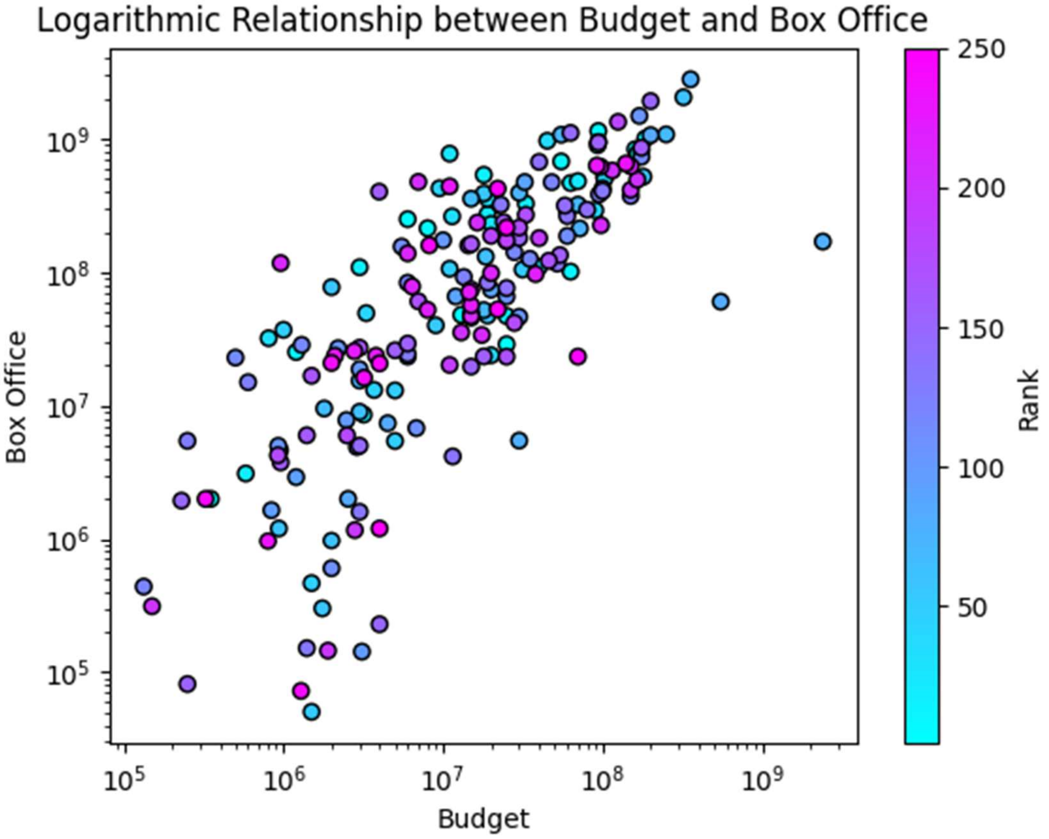


***Figure** 7 Relationship between budget and box office, alongside the rank*

The logarithmic relationship between budget and box office was of interest in Figure 1, so it was revisited in addition to colouration by rank. While a positive trend definitely exists, the budget nor the box office of a film appear to correlate directly to the rank it has. In order to confirm this, a 3D plot of the same variables was generated (Figure 8). By rotating the plot, it was observed that correlation between budget and box office indeed has no bearing on the rank of the movie, with low budget and high budget movies alike reaching both high and low into the ranks.

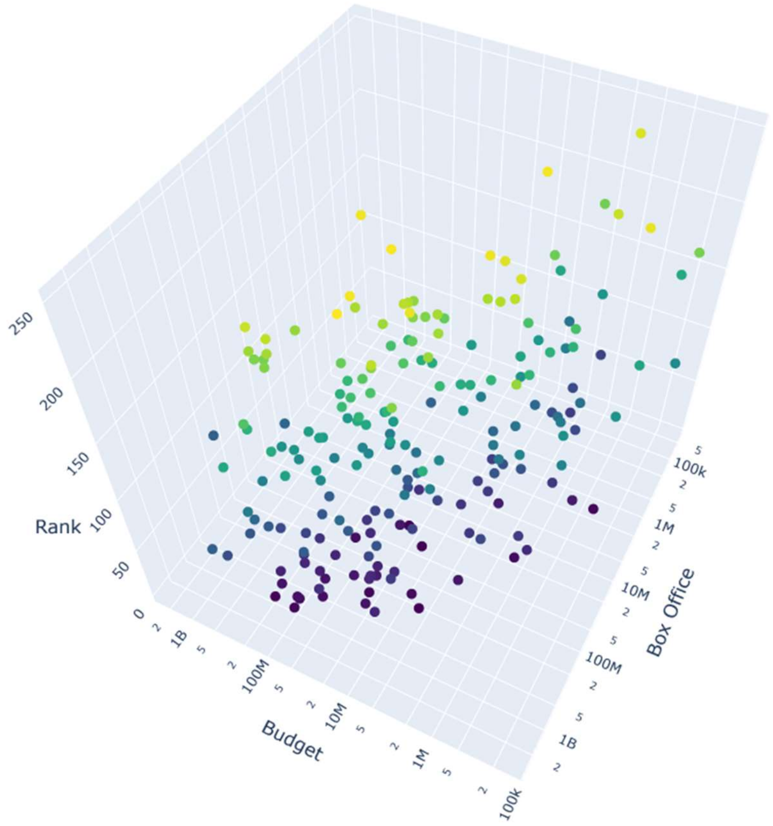


Figure 8a *Various rotations of a 3D implementation of Figure 7*

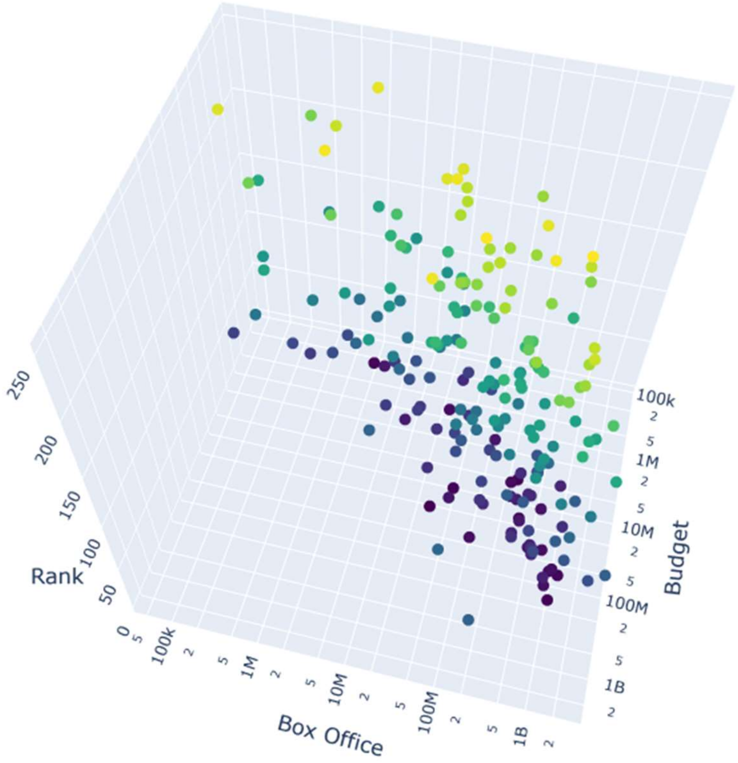


Figure 8b *Various rotations of a 3D implementation of Figure 7*

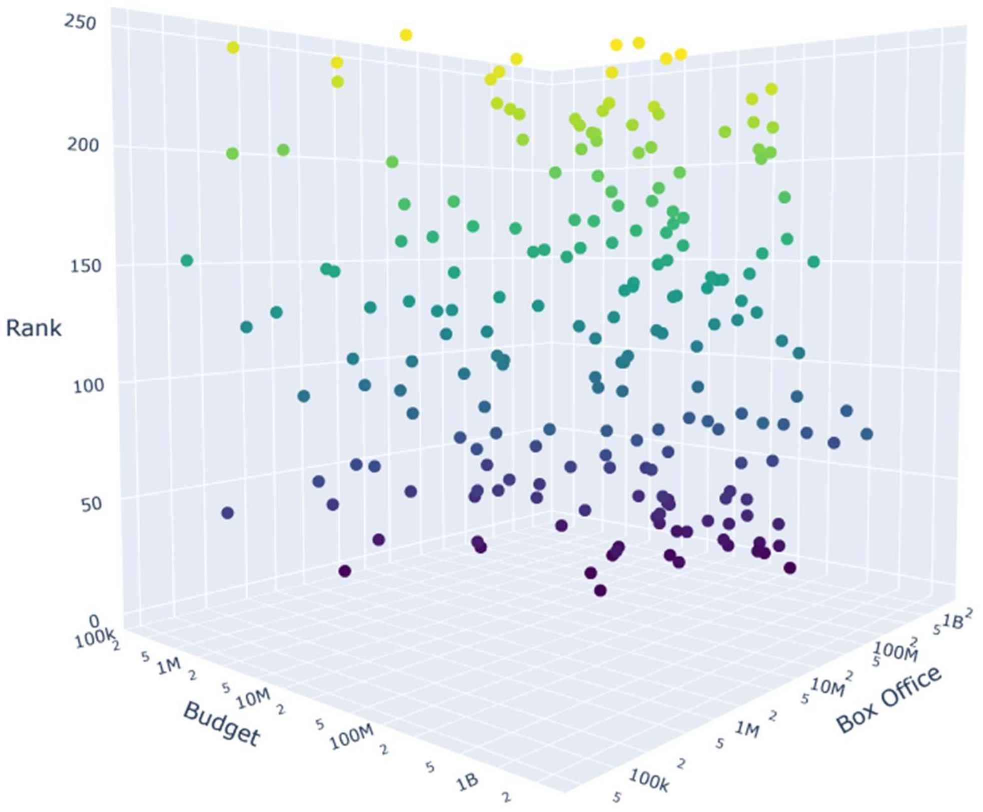


Figure 8c *Various rotations of a 3D implementation of Figure 7*

Self Evaluation

This project successfully used a multitude of data visualisation techniques to investigate several of the variable relationships underpinning the top 250 IMDb movies. To the best of my knowledge the data and visualisations are accurate as of the time the data was accessed. Although there were no groundbreaking results or hidden relationships between a movie's commercial success and any of the variables investigated, discovering trends that were readily explained by worldly phenomena such as the increase of animation films following the introduction of CGI made this project a personal success.

As for design choices, I was overall quite happy with the final visualisations I completed. Colour-wise, I ensured to the best of my ability that all variations of colour-blindness were accomodated for, excluding plots with the genres as the class due to the large number of colours required. I was particularly happy with how the heatmap turned out, and I think it perfectly illustrates an interesting relationship between the genres and their run times. Although I found it difficult to position the 3D scatterplot in a way that demonstrated the lack of trend, I was satisfied with the 3 perspectives I chose.

Despite these successes, there were several design choices that I recognise to be abnormal. For instance, the radar plot did not serve as valuable as I would have hoped; turning out convoluted and uninformative, despite my efforts to find meaningful relationships. In order to improve upon it, I would have split it into several subplots to prevent overcrowding. This overcrowding cost the graph its readability, and the presence of so many classes made colour choices difficult. In addition, I would have liked to investigate the normalisation further in an attempt to produce a graph with more defined distinctions between the genres.

Although the visualisations I completed were largely successful, I did not incorporate most of the non-numerical variables from the original dataset. Given the chance, I would like to create a web between directors and their casts to investigate whether certain duos appeared substantially more than others. Furthermore, I had initially planned on creating a wordcloud to determine if certain phrases or keywords in taglines and titles had impact on a movies commercial or public success. Unfortunately, I lacked the time and creativity to see this plan through.

If I had much more time to work on this project, I would have liked to find or collect data for a greater number of movies. In particular, the bottom 100 IMDb movies would have been an interesting contrast to the top 250. IMDb appears to have limited certain attributes such as budget and box office recently, so I would have to source that data elsewhere.

Overall, I think I deserve a 5-6 for this project. Time management was a big struggle throughout, and I'm convinced there were several other visualisations I could have done if I had spent more time on the project. Furthermore, none of the visualisations were particularly novel or creative, so implementing the aforementioned wordcloud would have helped in that regard. Altogether, I still learned a lot and am glad I had this experience.

Conclusion

In conclusion, although none of the variables investigated in this report had a remarkable effect on the ranking of the film, there were still several interesting trends to be observed between the other variables. With a larger dataset and films with greater variance in their ratings, more interesting relationships should be observed; or at least strengthen existing ones found in this report. Filmmaking very much remains a passion first and foremost for many directors, and although techniques both in cinematography and marketing will continue to evolve, it is unlikely we will discover the formula to a perfect movie any time soon.

References

Raju, C. (2023). IMDb Top 250 Movies Dataset. Retrieved from Kaggle: https://www.kaggle.com/datasets/rajugc/imdb-top-250-movies-dataset

Dow, D. (2009). Motion Picture Ratings. Retrieved from The First Amendment Encyclopedia: https://www.mtsu.edu/first-amendment/article/1247/motion-picture-ratings#:~:text=or%20ban%20movies.-,In%201968%20the%20Motion%20Picture%20Association%20of%20America%20(MPAA)%20established,submit%20their%20films%20for%20rating

Formichella, L. (2020). 14 groundbreaking movies that took special effects to new levels. Retrieved from Insider: https://www.insider.com/most-groundbreaking-cgi-movies-ever-created-2020-1

Choi, D. (2021). Guns, Muscles, and Kung Fu – The 1980s and the Birth of the “Action Movie” and “Action Hero”. Retrieved from Hollywood Insider: https://www.hollywoodinsider.com/1980s-birth-of-action-movie/