IMDB Movie Analysis

Project Description:

The **IMDB Movie Analysis** project aims to uncover the key factors that influence a movie's success on IMDB, where success is defined by high ratings. By analyzing various attributes such as genre, director, cast, budget, and release date, this project seeks to provide valuable insights for movie producers, directors, and investors. Through data collection, cleaning, exploratory analysis, and predictive modeling, the project will identify significant trends and develop tools to forecast the potential success of future movies. Ultimately, this analysis will help industry stakeholders make informed, data-driven decisions to enhance the success rates of their projects.

Approach:

Data pre-processing, cleaning and error rectification:

- Out of 5043 (not counting the headers row) rows 126 rows with duplicate data (filtered by movie name) were removed. Rows with 8 or more empty column entries were deleted. 4909 rows remained. Out of these rows, three rows containing video games were removed. We are now left with 4906 rows.

- Columns like 'color', 'aspect ratio', 'facenumber_in_poster' and 'plot keywords' were deleted as those didn't seem important.

- In these rows blank spaces in columns, 'director name', 'duration', actor_1_name, actor_2_name, actor_3_name, 'num_user_for_reviews', 'language', 'country' and 'content_rating' were filled with correct values after referring to their websites.

- Blank values in budget, gross and 'director_facebook_likes' were replaced with correct values if available or replaced with NA.

- Data was cleaned by removing random characters or replacing with correct letters in the director and movie columns.

- Some entries in the country column were given as "Official Site". They were replaced with the correct country name.

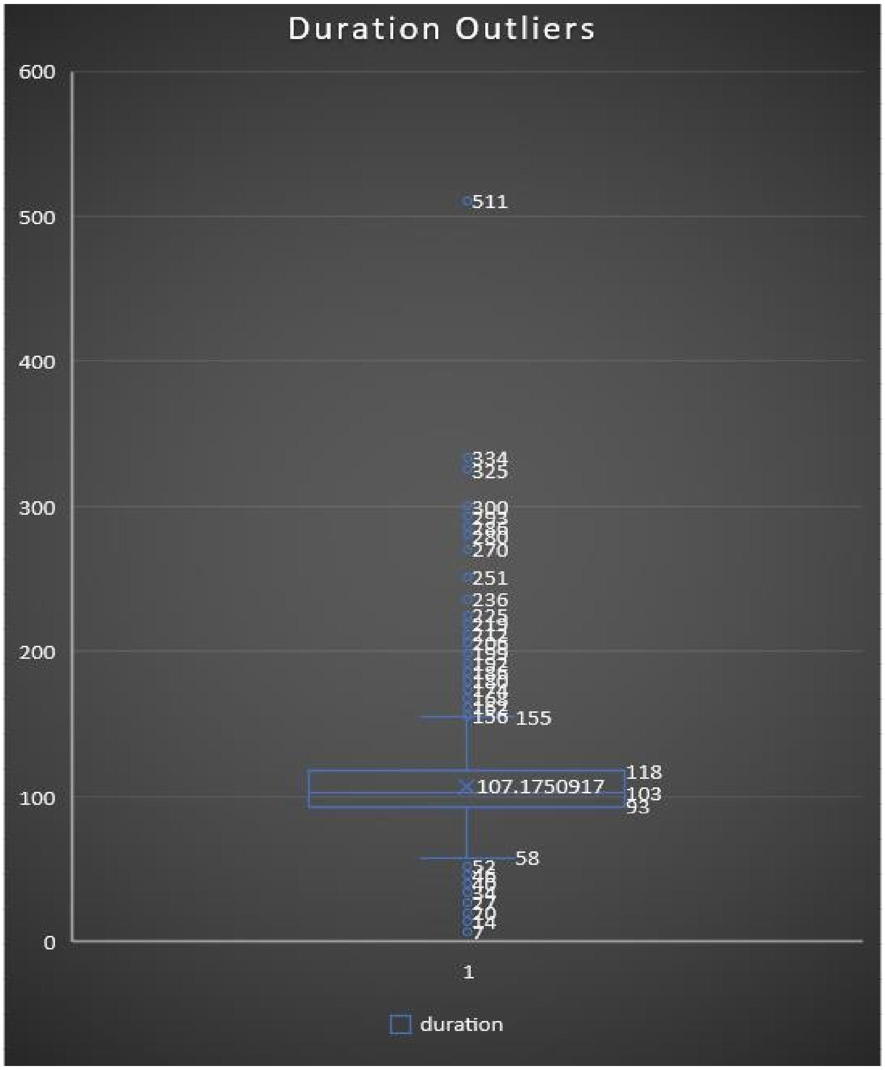
- The entries in language column “None” were also replaced with right values.
- Budget values of some movies were given in currencies other than dollars. Those were converted into dollars.
- Outliers were recognized for duration column and replaced with median values.
- A new column was added “no. of genres” signifying the number of genres shown in genre column.
- Replaced “Musical” genre with “Music” genre as it means the same.
- “West Germany” in country column was replaced with “Germany”.

Tech-Stack Used:

Microsoft Excel 365:

This tech stack was used as it is the most convenient and efficient software to use to process data in a tabular format, make pivot tables and present data by visualizations. The hyperlink for excel sheet in which operations are performed is here:

https://docs.google.com/spreadsheets/d/1kYbc6py20C4XbfnvGzUpLbdeie6DUEpq/edit?usp=drive_link&ouid=110446323541757318313&rtpof=true&sd=true



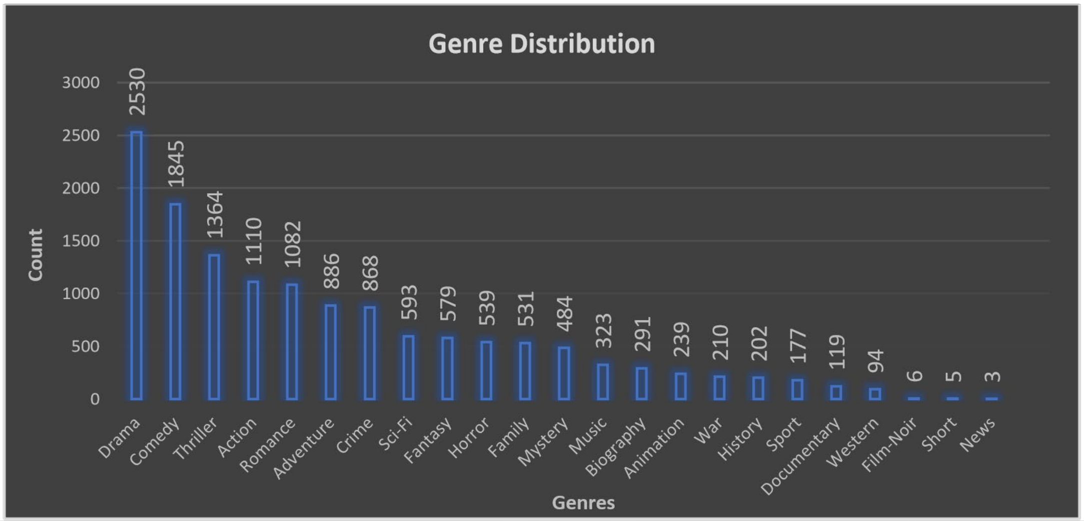
Insights:

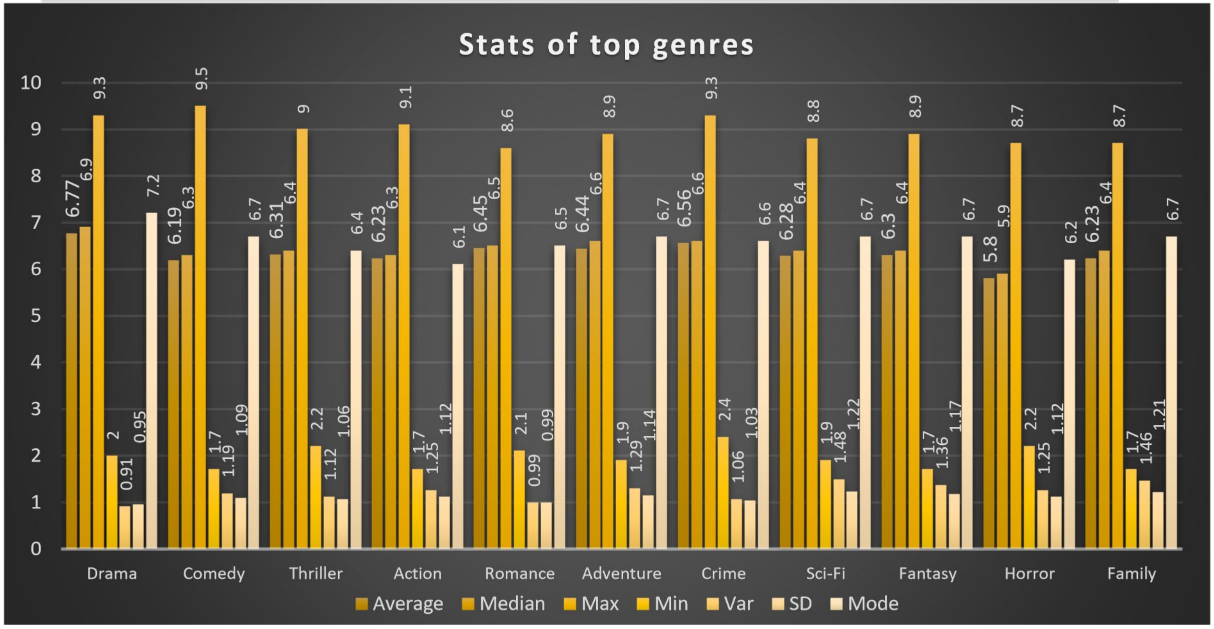
A. Movie Genre Analysis

Analyze the distribution of movie genres and their impact on the IMDB score.

- **Task:** Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.

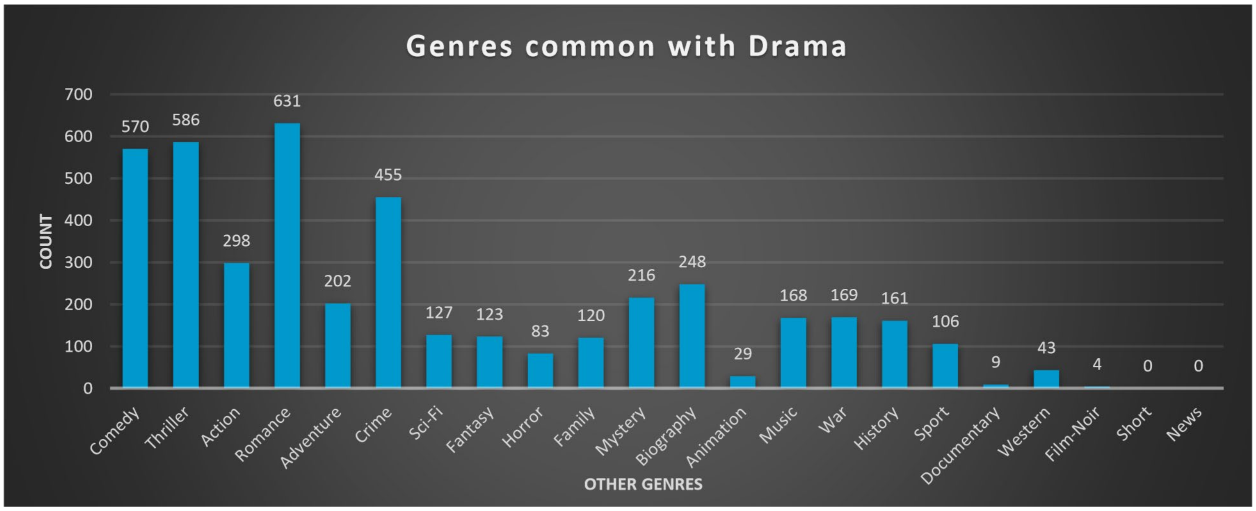
- **Result:** Following is the count of each genre. Some movies have a single genre and some have multiple genres. The top genres are: Drama, Comedy, Thriller, Action, Romance, Adventure, Crime, Sci-Fi, Fantasy, Horror and Family





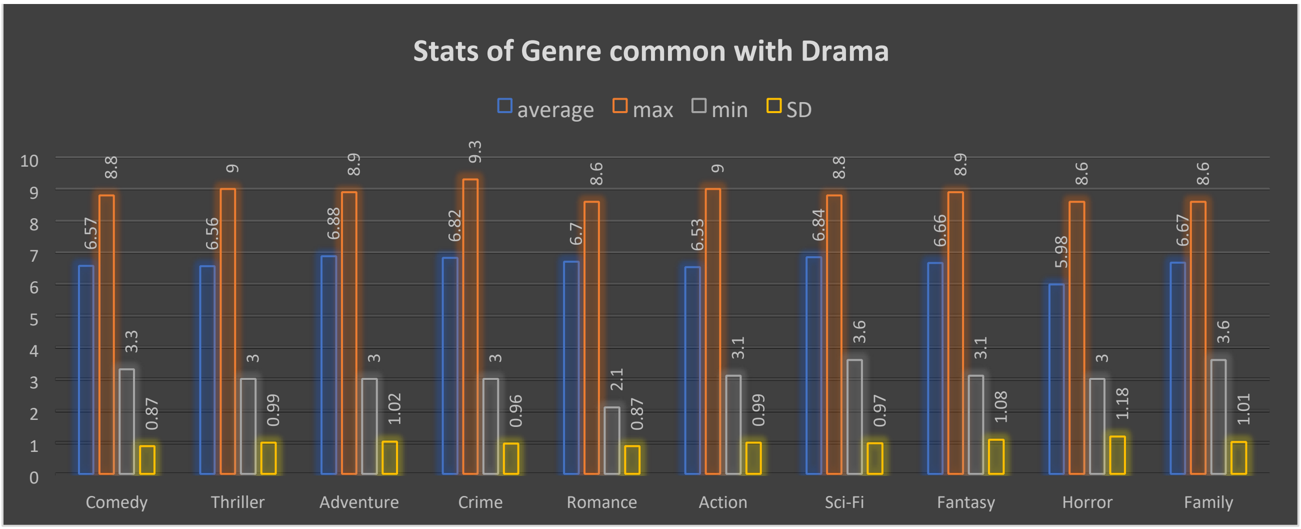
Q1. Why does Drama have the highest count in genre?

It may be due to the fact that most people want movies to invoke strong emotions in them which leads to the audience feeling a connection to the movie. Also, drama to a large degree, entails some other forms of genres or is used as a secondary genre to other genres. This may help in attracting a larger audience.



Q2. Why do a lot of movies with drama genre have other genres in combination with Drama?

Drama genre has the highest count of movies however it's mostly in conjunction with other genres and not on its own. These other genres in combination with the Drama genre can give an idea about the kind of movie scripts or movie themes that the audience is more engaged in. According to the following charts, Comedy, Romance, Thriller and Crime are the genres used most in combination with the Drama genre.



Q3. Why are Comedy, Romance, Thriller and Crime the genres used most in combination with the Drama genre?

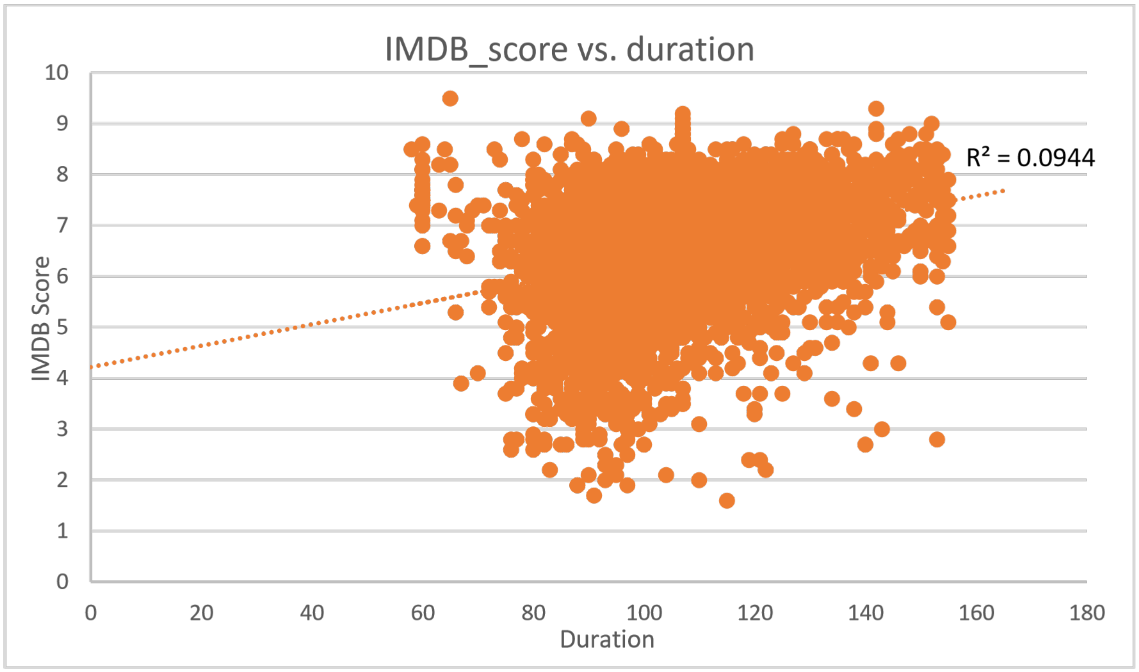
These genres in association with the drama genre, create stories and movie scripts that invoke a variety of emotions in the audiences. They reflect common themes and scenarios that are relatable to a wide audience. Attracting a larger audience ensures higher viewership and commercial success. While the rest of the genres might invoke stronger responses from a niche group of people but don't attract an audience universally.

B. Movie Duration Analysis:

Analyze the distribution of movie durations and its impact on the IMDB score.

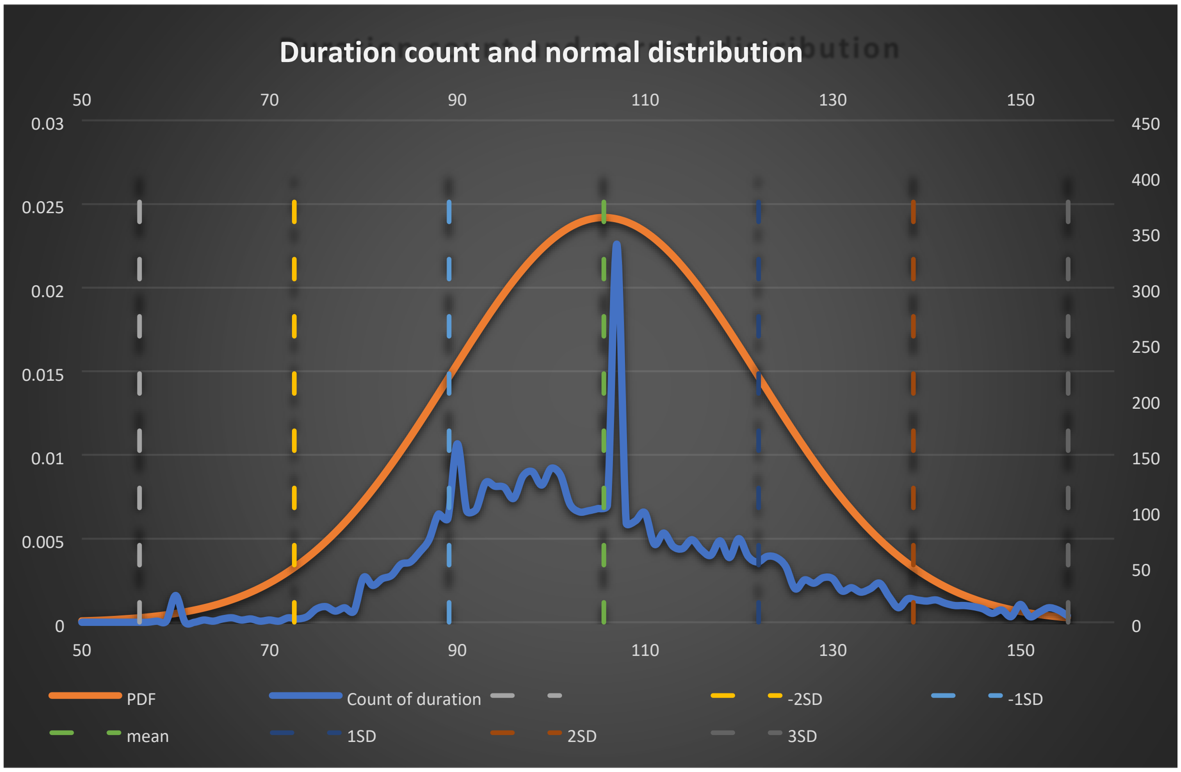
- **Task:** Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

- **Result:** Following plot displays the relationship between movie duration and IMDB score. The Rsquared value represents how strong the relationship between two variables is. A value closer to 1 signifies a strong relationship. The trendline seems to suggest a positive correlation between duration and IMDB scores.



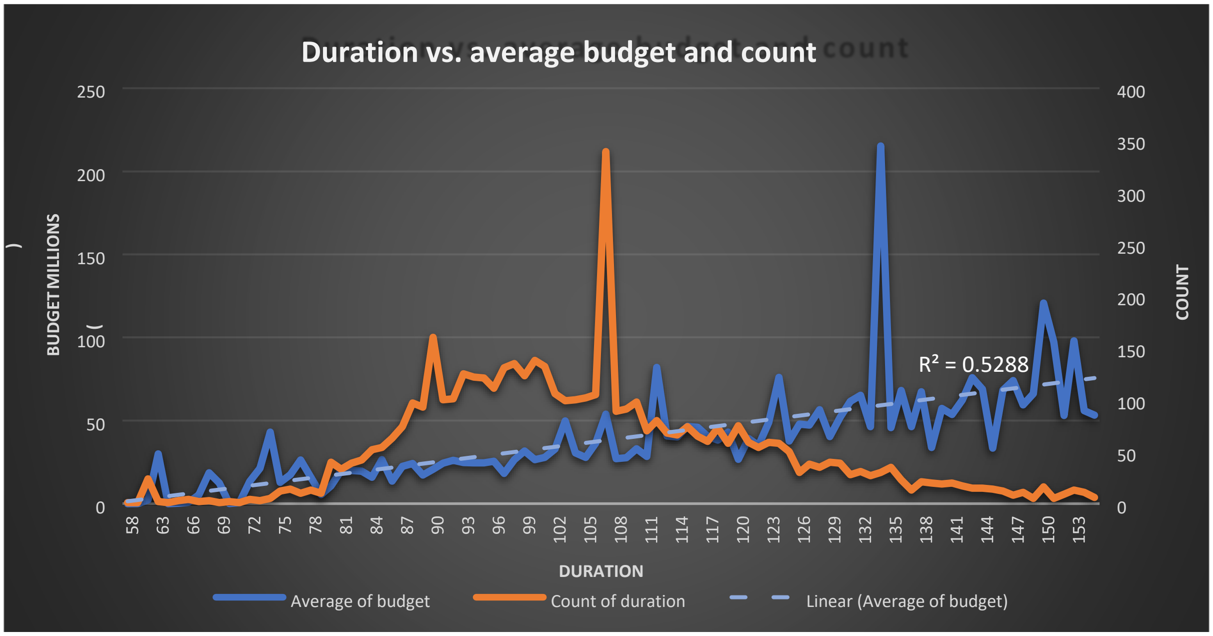
Q1. Why is the correlation between duration and IMDB scores weak?

The low R-squared value indicates no strong relationship between duration and IMDB scores. This may be due to the fact that most datapoints lie between 80 to 120 minutes and the sample points are less towards either ends. The following graph shows a normal distribution for duration, count and IMDB scores. Therefore, 67% of datapoints lie between 90 to 120 minutes.



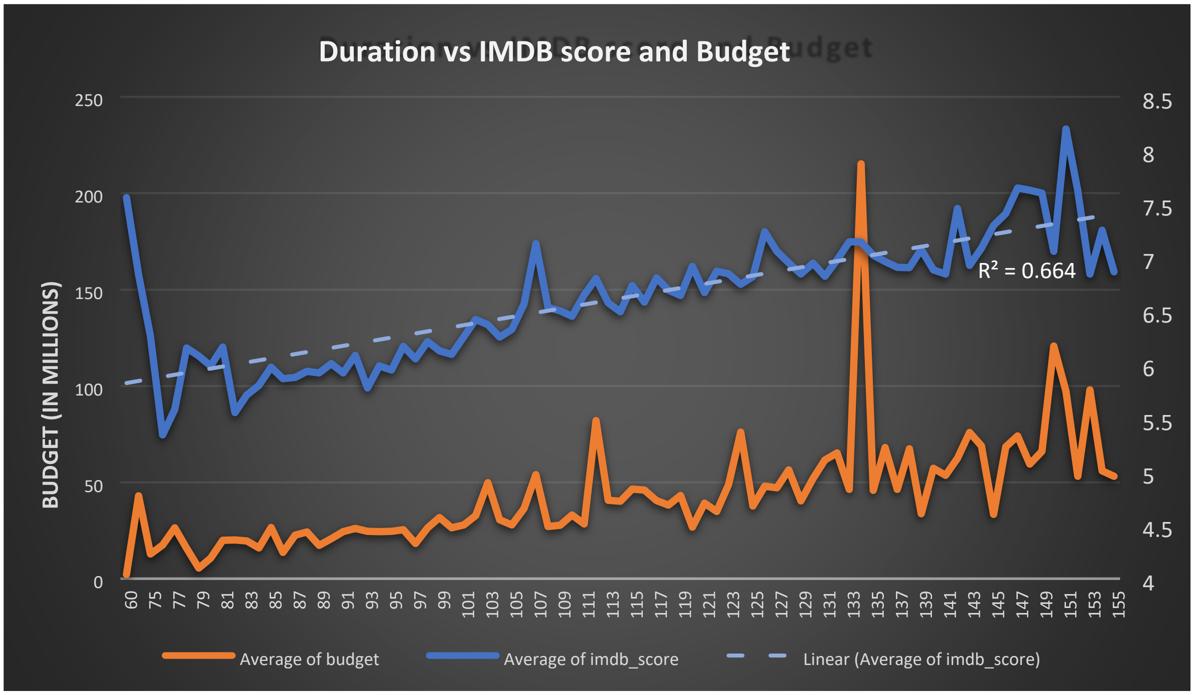
Q2. Why do most movies have a duration between 90 to 120 minutes?

According to the following chart, most movies of a higher duration require a higher budget. There is a moderate correlation between duration of a movie with its budget suggesting a positive correlation. It makes sense that the creators may want to balance movie duration with budget for production.



Q3. Why do movies with higher duration require higher budget?

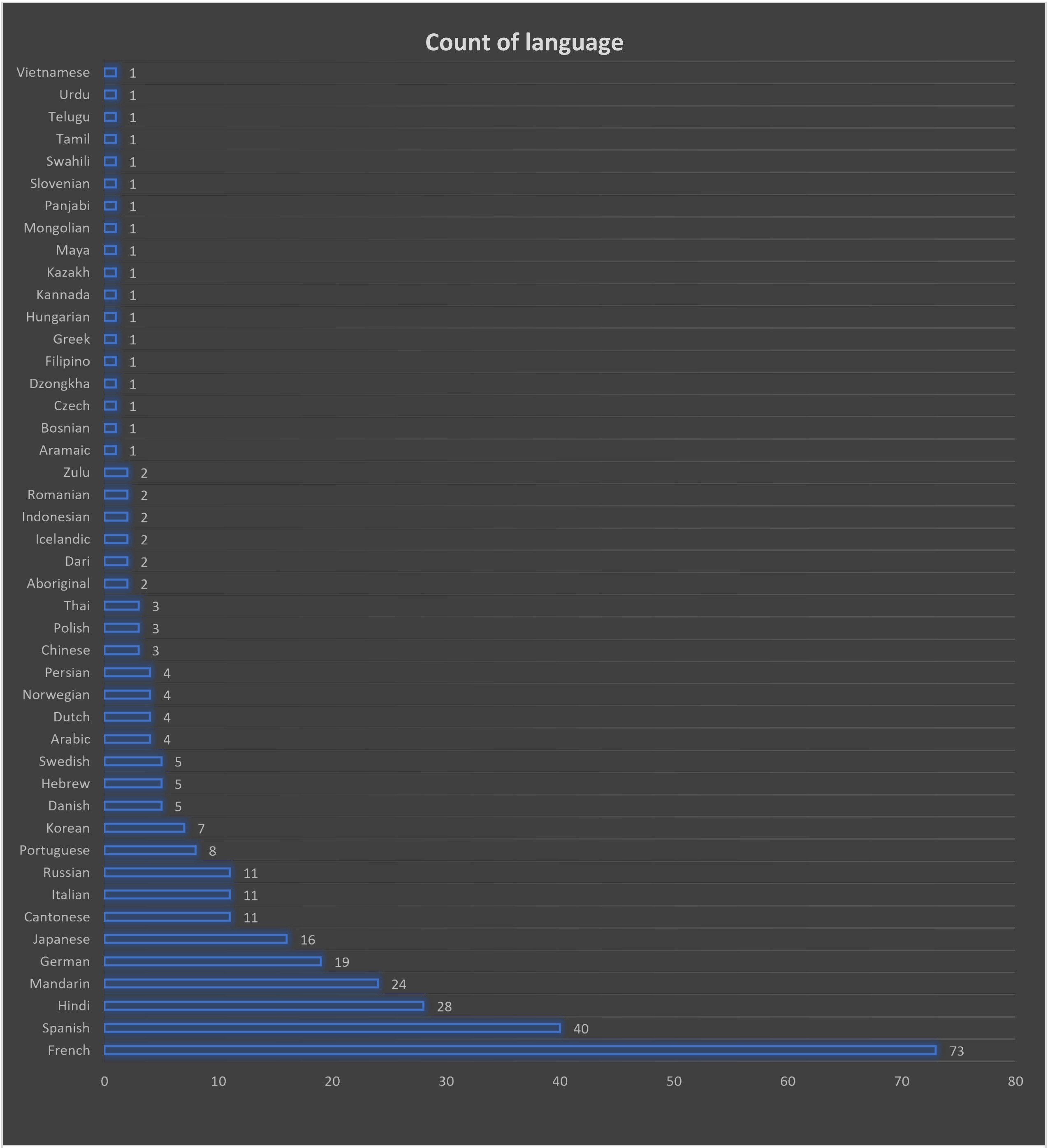
This may be understood by the following chart. There is a positive correlation between movie durations and the average IMDB scores. Also, the average budget increases along with duration of the show. This may be due to the fact that higher budget leads to higher production quality and better resources which enhances audience's experience leading to better ratings. Since there is a moderate relationship between duration and budget, the same can be said about duration and IMDB scores.

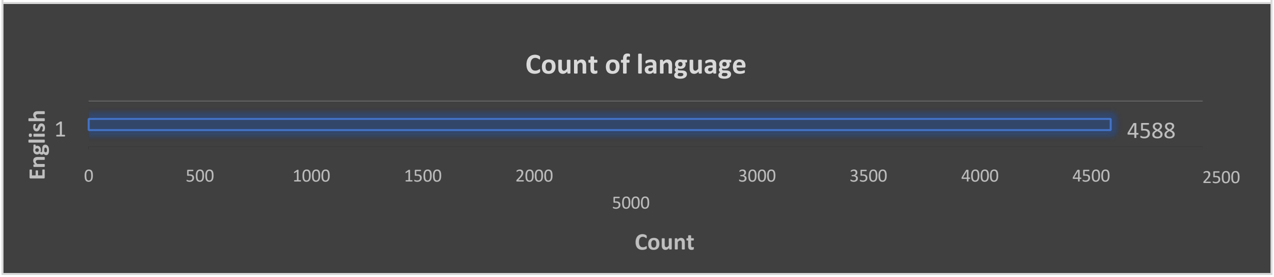


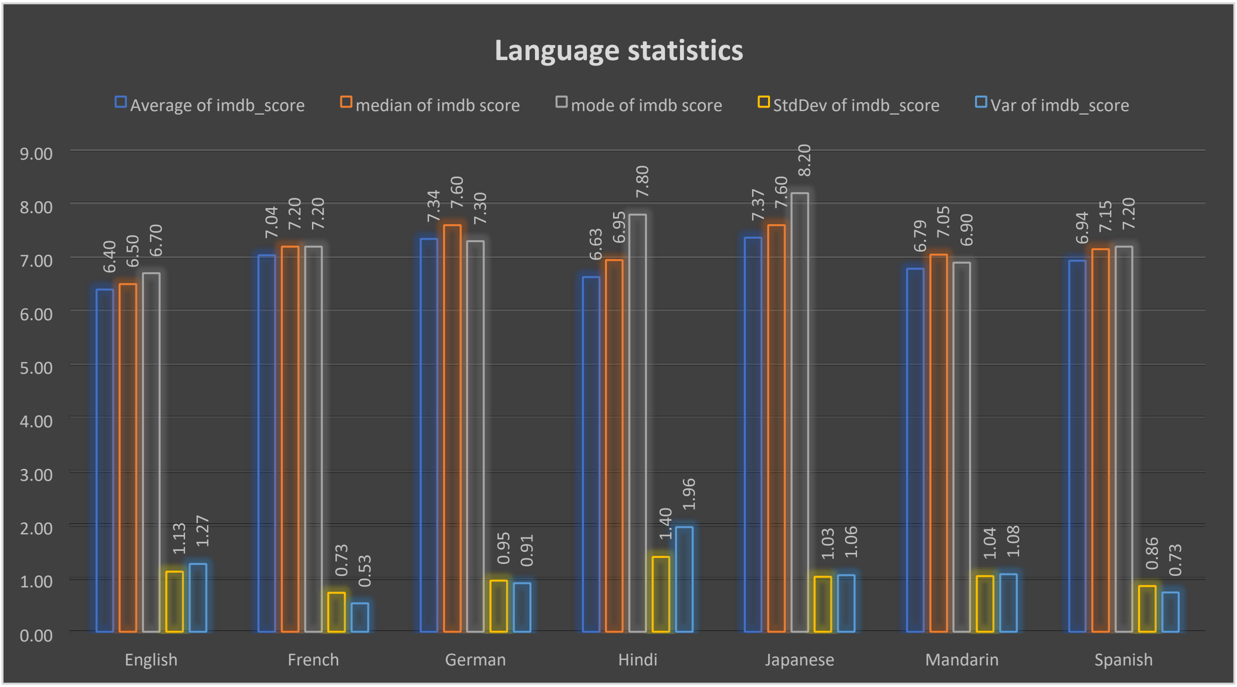
C. Language Analysis:

Situation: Examine the distribution of movies based on their language.

- **Task:** Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.
- **Results:** The following graph shows that English language dominates the list of movies, followed by French, Spanish, Hindi, Mandarin, German and Japanese.

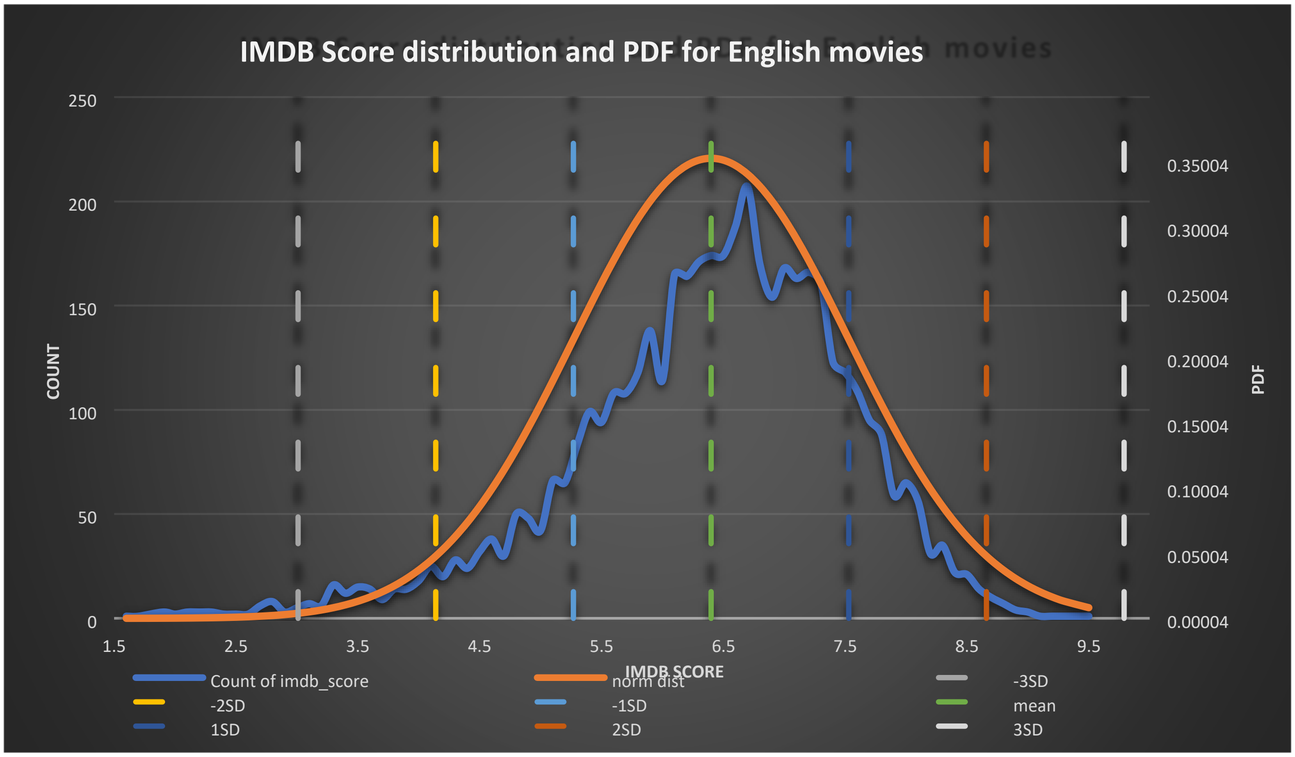

Count of language
Vietnamese 1
Urdu 1
Telugu 1
Tamil 1
Swahili 1
Slovenian 1
Panjabi 1
Mongolian 1
Maya 1
Kazakh 1
Kannada 1
Hungarian 1
Greek 1
Filipino 1
Dzongkha 1
Czech 1
Bosnian 1
Aramaic 1
Zulu 2
Romanian 2
Indonesian 2
Icelandic 2
Dari 2
Aboriginal 2
Thai 3
Polish 3
Chinese 3
Persian 4
Norwegian 4
Dutch 4
Arabic 4
Swedish 5
Hebrew 5
Danish 5
Korean 7
Portuguese 8
Russian 11
Italian 11
Cantonese 11
Japanese 16
German 19
Mandarin 24
Hindi 28
Spanish 40
French 73
0
10
20
30
40
50
60
70
80


Count of language
English
1
4588
0
500
1000
1500
2000
3000
3500
4000
4500
2500
5000
Count



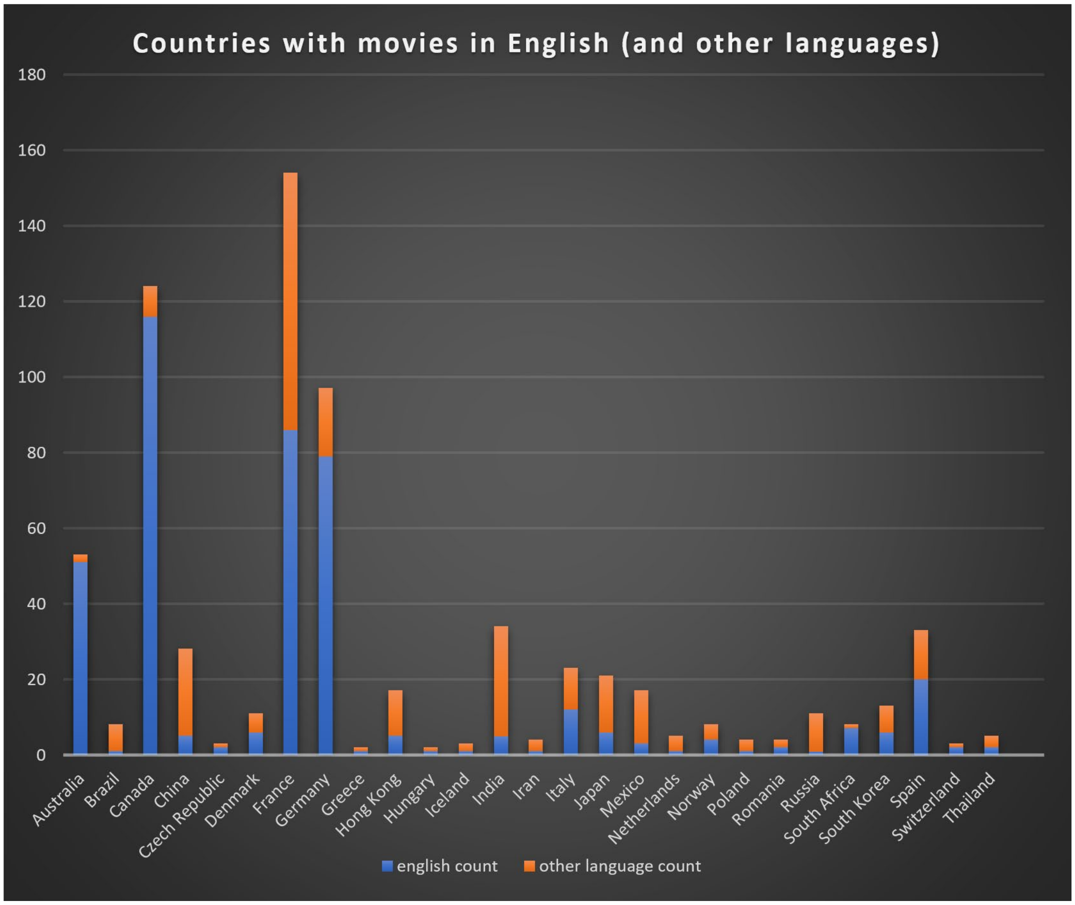
Q1. Why does English have overall the lowest median and comparatively higher standard deviation?

English has the highest datapoints which also follow the normal distribution curve. Most movies (around 67%) fall into the IMDB Score range 5.5-7.5 which lowers the median IMDB Score. The presence of datapoints on extreme ends of the scale contributes to a slightly higher standard deviation.



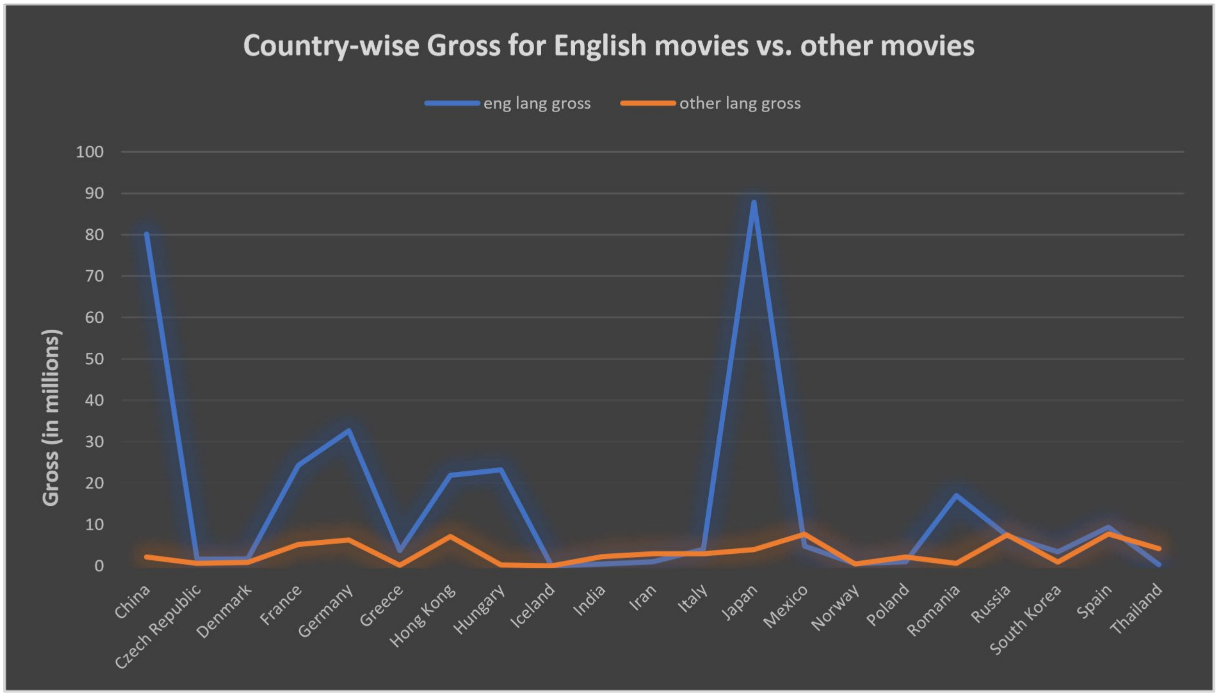
Q2. Why is English the most popular language for movies?

English is the most popular language spoken across the world. Coupling that with the success of Hollywood, it makes sense that most movies on the list are produced in English as USA along with a few other countries is an English-speaking country. A few other countries have also produced English movies.



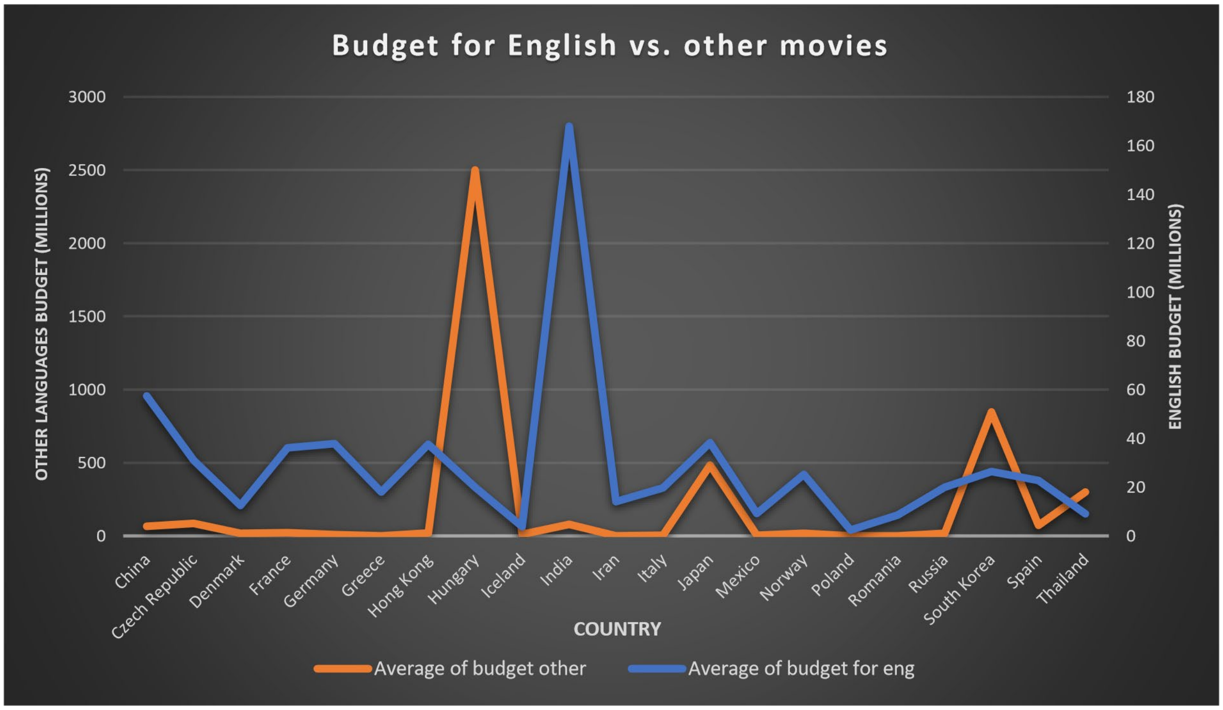
Q3. Why do countries that don’t speak English as their primary language make English movies?

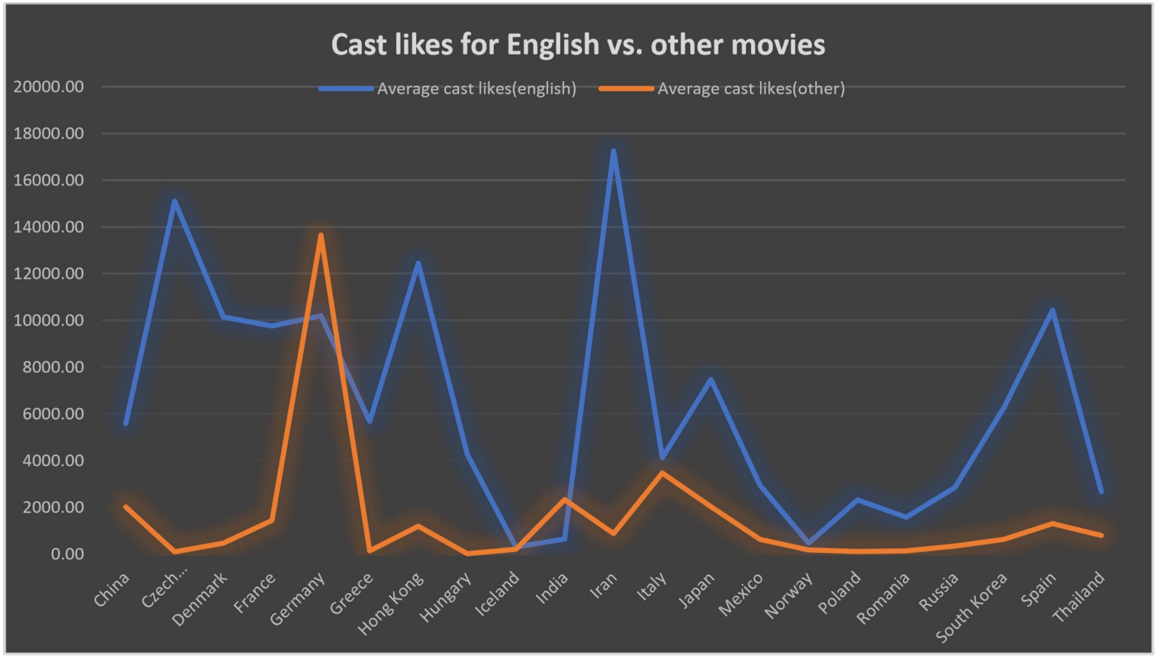
According to the following chart, for most countries, the average gross made by English movies is higher than the average gross made by other movies. Another reason could be Hollywood directors or production members taking an interest in other countries and collaborating with those to produce English movies as special projects to exploit the interests of international fans as well as local fans.



Q4. Why is the gross revenue collected higher for English movies than gross revenues of movies in regional languages?

According to the following charts, even though the budget for most movies is similar regardless of the language (except for Hungary, Japan, India and South Korea), the gross revenue is higher for English movies. This could be due to the production casting popular actors for their cast in English movies as opposed to regular movies. This may have contributed to audience having a greater response to movies produced in English.

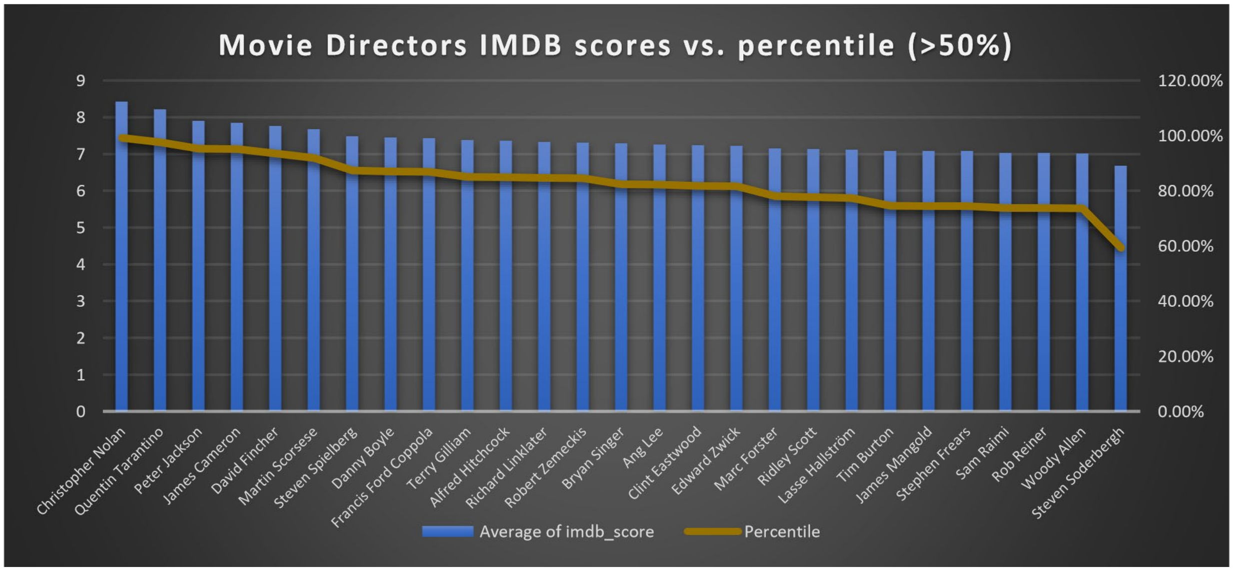




D. Director Analysis:

Influence of directors on movie ratings.

- **Task:** Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.
- **Result:** Top performing directors were determined by selecting directors with percentile >50% in IMDB scores. These were further narrowed down by only selecting directors who have produced 8 or more movies, percentile above 70%, range less than or equal to 3 and average IMDB score of 7. This was done to verify and only keep directors that have consistently created bestperforming movies. Finally, ten directors were determined to be the top performing directors.

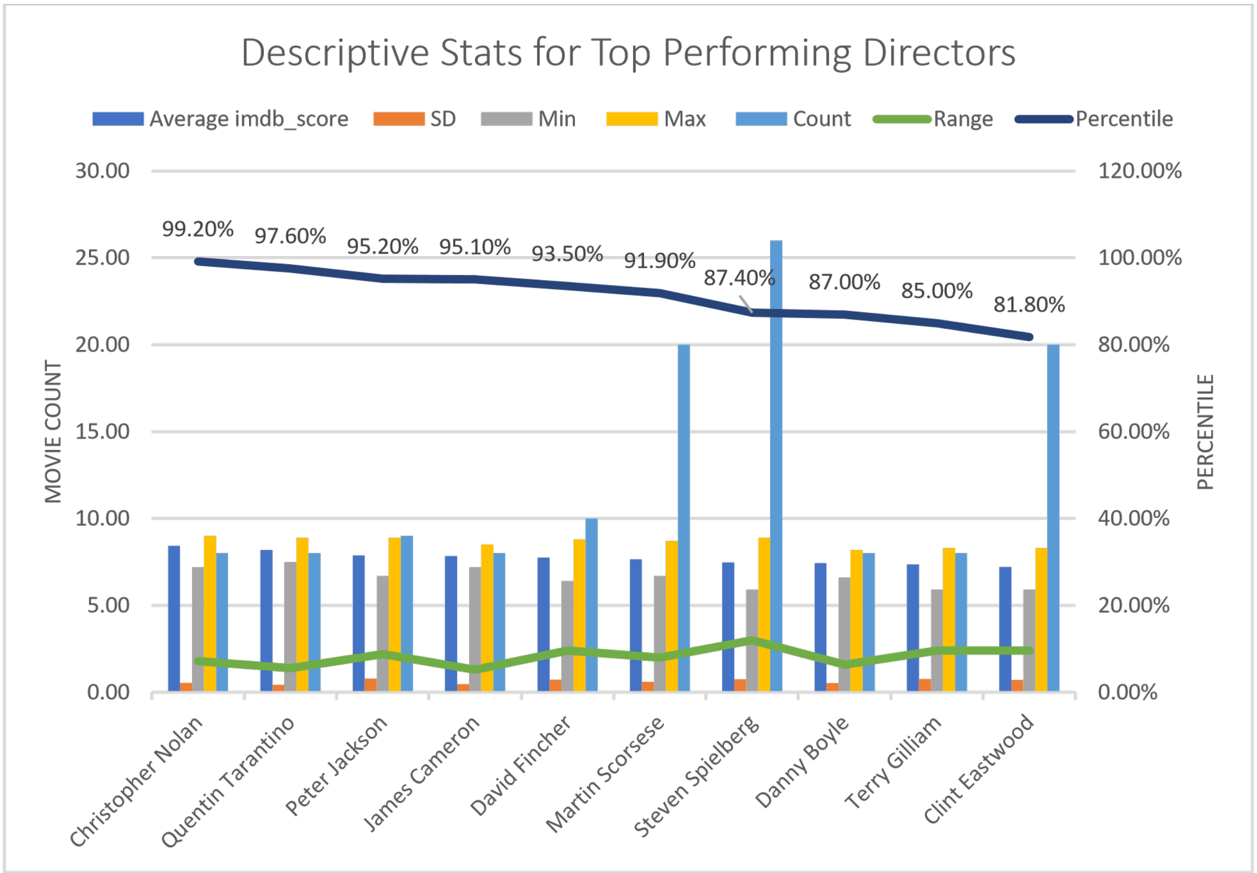


Movie Directors IMDB scores vs. percentile (>50%)

Top performing directors	Average IMDB score	SD	Min	Max	Count	Range
Christopher Nolan	8.43	0.54	7.2	9	8	1.8
Quentin Tarantino	8.20	0.42	7.5	8.9	8	1.4
Peter Jackson	7.89	0.77	6.7	8.9	9	2.2
James Cameron	7.85	0.46	7.2	8.5	8	1.3
David Fincher	7.75	0.72	6.4	8.8	10	2.4
Martin Scorsese	7.66	0.60	6.7	8.7	20	2
Steven Spielberg	7.48	0.74	5.9	8.9	26	3
Danny Boyle	7.44	0.52	6.6	8.2	8	1.6
Terry Gilliam	7.36	0.77	5.9	8.3	8	2.4
Clint Eastwood	7.23	0.70	5.9	8.3	20	2.4

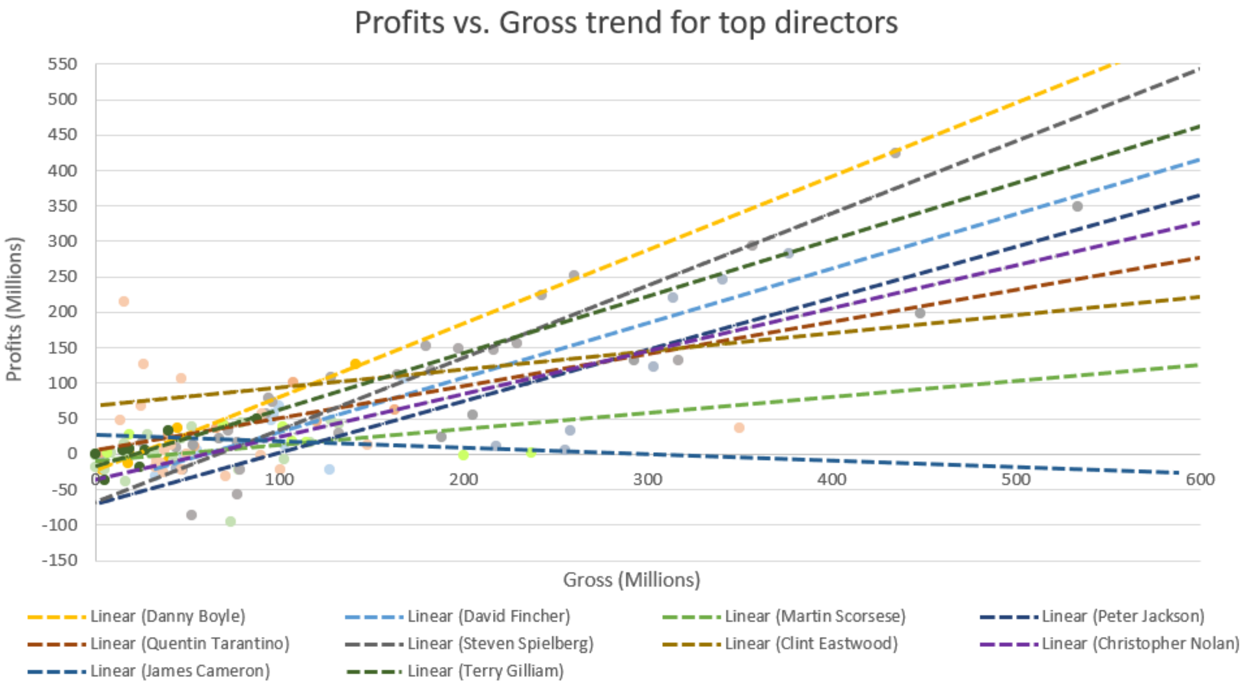
Q1. Why are these directors the top performing directors?

The above mentioned 10 directors all have the average IMDB score in the >80% range. They all have also produced a minimum of 8 movies and have low standard deviations for the IMDB scores of their movies. The directors have directed movies averaging on an IMDB score of 7 or above and have a low range (good consistency).



Q2. Why is it important to determine top-performing directors?

A consistently high record of IMDB scores contributes to stronger statistics of these directors. These stronger statistics is a result of positive response from the audience leading to more engagement and a higher box-office revenue. The following chart shows trendlines for gross revenue vs. profits made by the movies of mentioned directors. Judging by the R-squared values, for most directors (except two) these directors push out movies that almost certainly increase in margin with increase in gross revenue.



Top performing directors	R-squared values
Christopher Nolan	0.8885
Quentin Tarantino	0.5498
Peter Jackson	0.6086
David Fincher	0.512
Martin Scorsese	0.0899
Steven Spielberg	0.8405
Danny Boyle	0.9145
Terry Gilliam	0.6581

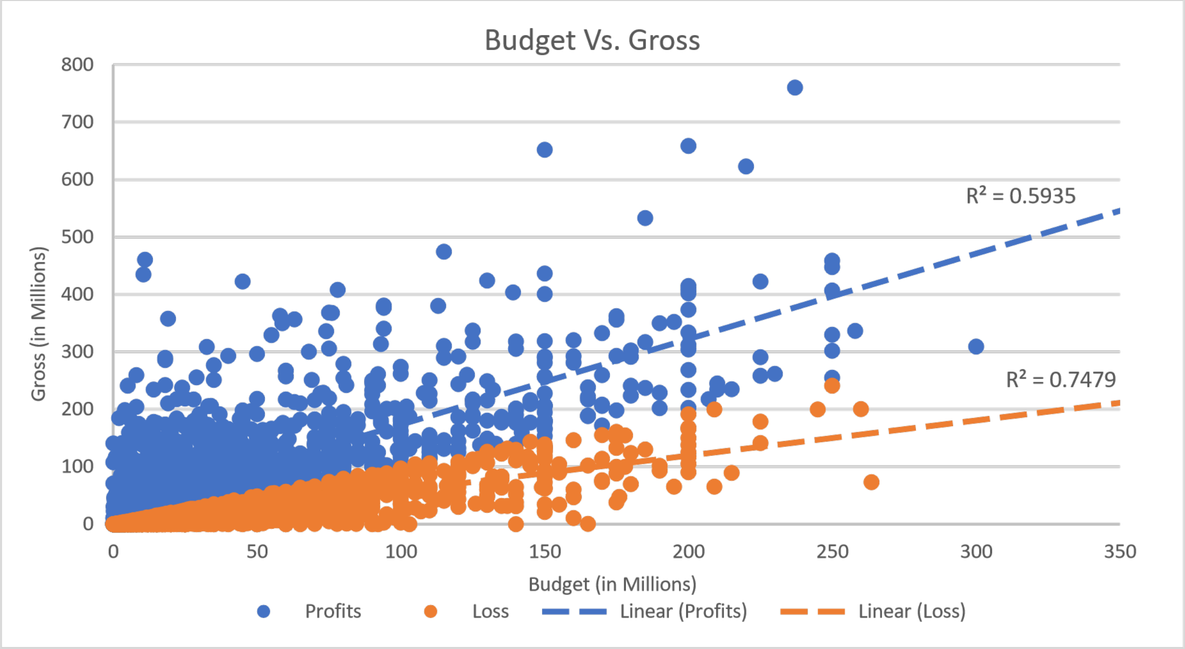
Q3. Why do these directors have movies that almost certainly increase in margin with increase in gross revenue?

The production team’s effective budget management may play a crucial role. By wisely allocating resources, they can hire better cast members and optimize other aspects of production, which further enhances the movie’s appeal and profitability.

E. Budget Analysis:

Explore the relationship between movie budgets and their financial success.

- **Task:** Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
- **Result:** A comparison between movie budgets vs. movie gross revenue and movie budgets vs. profits made shows that most movies were profitable in number. However, the R-squared value was higher for movies that incurred a loss than for those that generated a profit. Finding answers for why most movies make a profit as well as why most movies are likely to incur a loss may give some helpful insights.

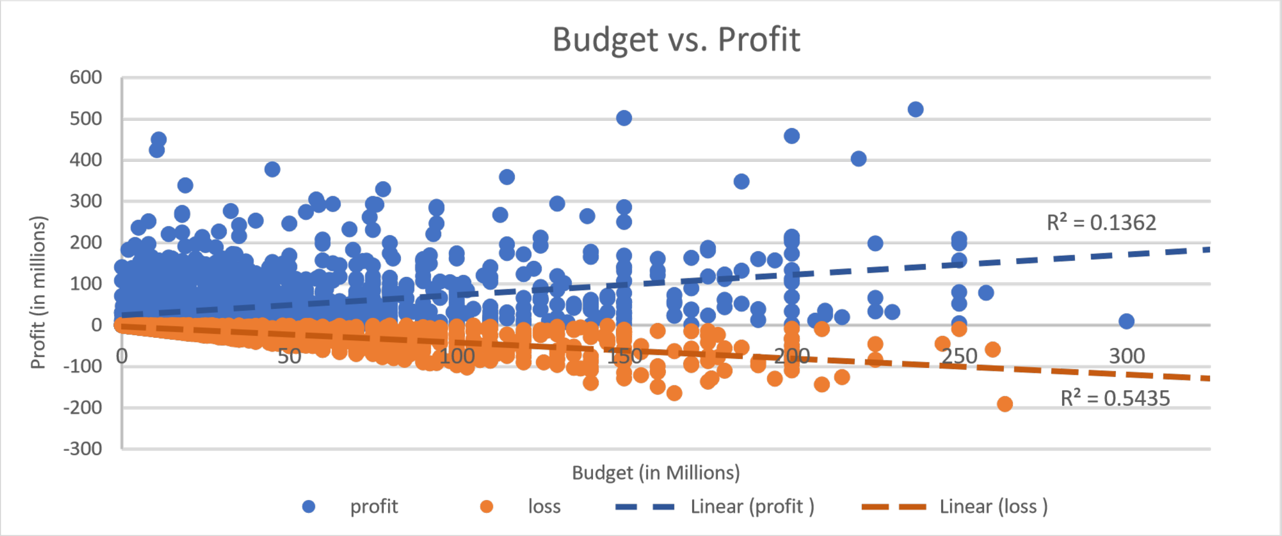


Q1. Why are the R-squared values bigger for movies that incurred a loss?

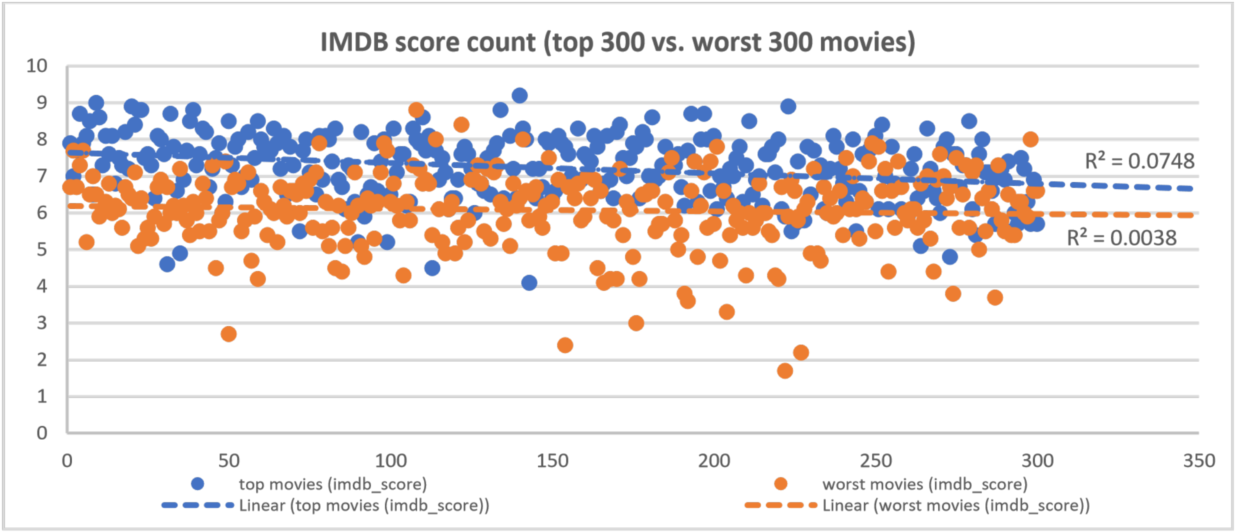
A bigger R-squared value means that the higher the budget of a movie is, the more likely it is to incur a loss instead of making a profit. However, R-squared values for both profitable movies and loss-making movies are more than 0.5 in the first chart indicating a strong relationship for both scenarios. Therefore, a higher budget does not entirely guarantee a higher gross revenue. More analysis is required to determine what makes a movie a financial success.

Q2. Why does higher budget not guarantee higher profits?

A chart for correlation of budget vs. profits shows a moderate relationship between loss incurred and budget meanwhile a weaker relationship for budget vs. profitable movies. Further investigations into budget and other factors for both types of movies may give more insights. These other factors may include movie genres, IMDB scores, cast, directors, etc. These relationships can be studied by taking a few movies that were the most profitable and few movies that incurred the most loss.



Elements like genre, cast, director, and marketing can significantly impact a movie's appeal and performance. The following charts show the correlation between the most and least profitable Movies and the IMDB scores. There is almost no correlation between IMDB scores and the top and worst movies depending on their profitability. However, most profitable movies have the IMDB score between 6 and 9.



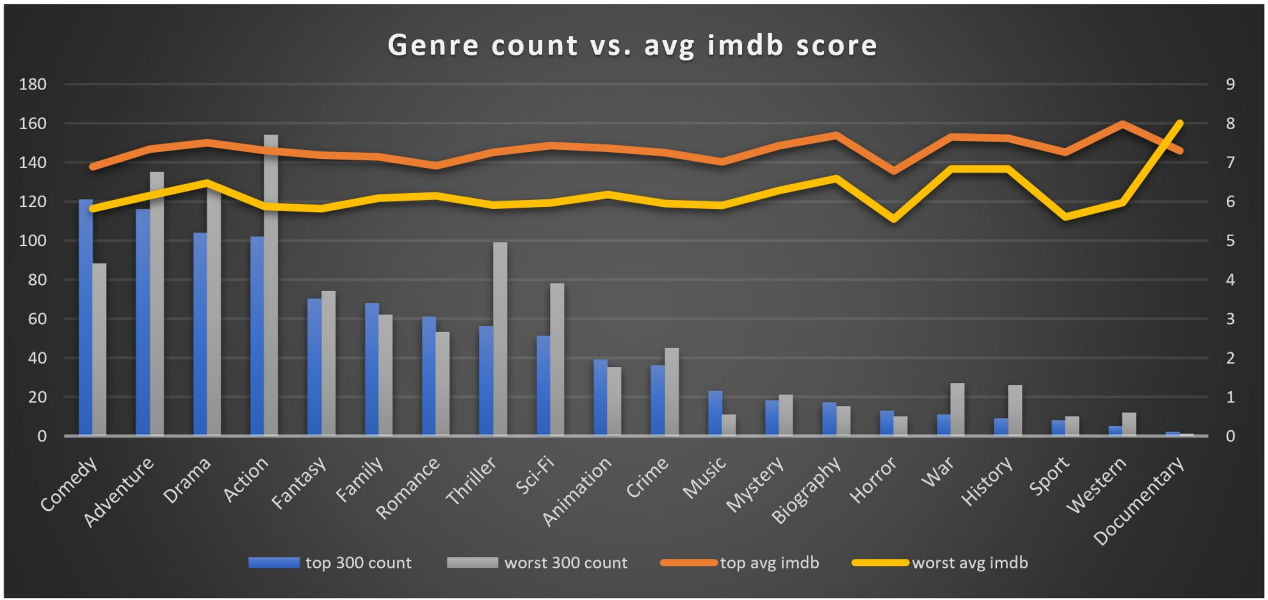
Q3. Why is there almost no correlation between IMDB scores and the profitability of movies?

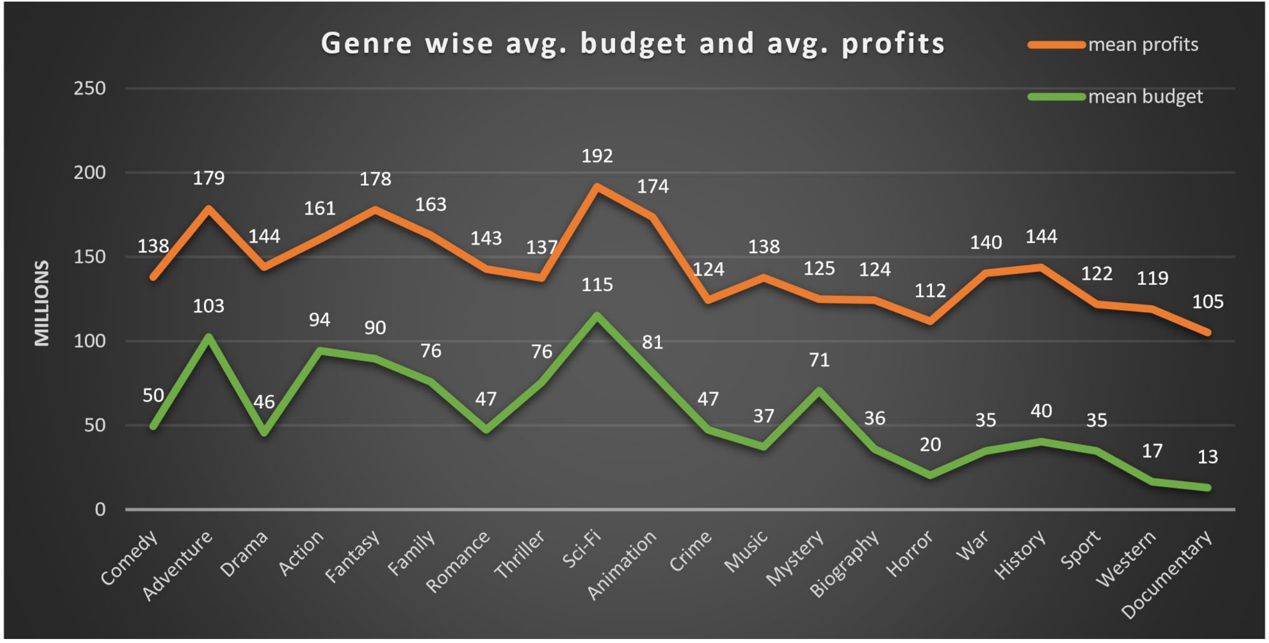
Even though most profitable movies have IMDB scores between 6 and 9, the correlation is very weak. This suggests that while having an above-average IMDB score might be beneficial, it is not the sole determinant of a movie's financial success. Other factors such as genre, marketing strategies, promotions, choice of cast, and directors play significant roles. Therefore, it may be more insightful to analyze the correlation between genre, budget/profit, and average IMDB scores to understand the broader picture.

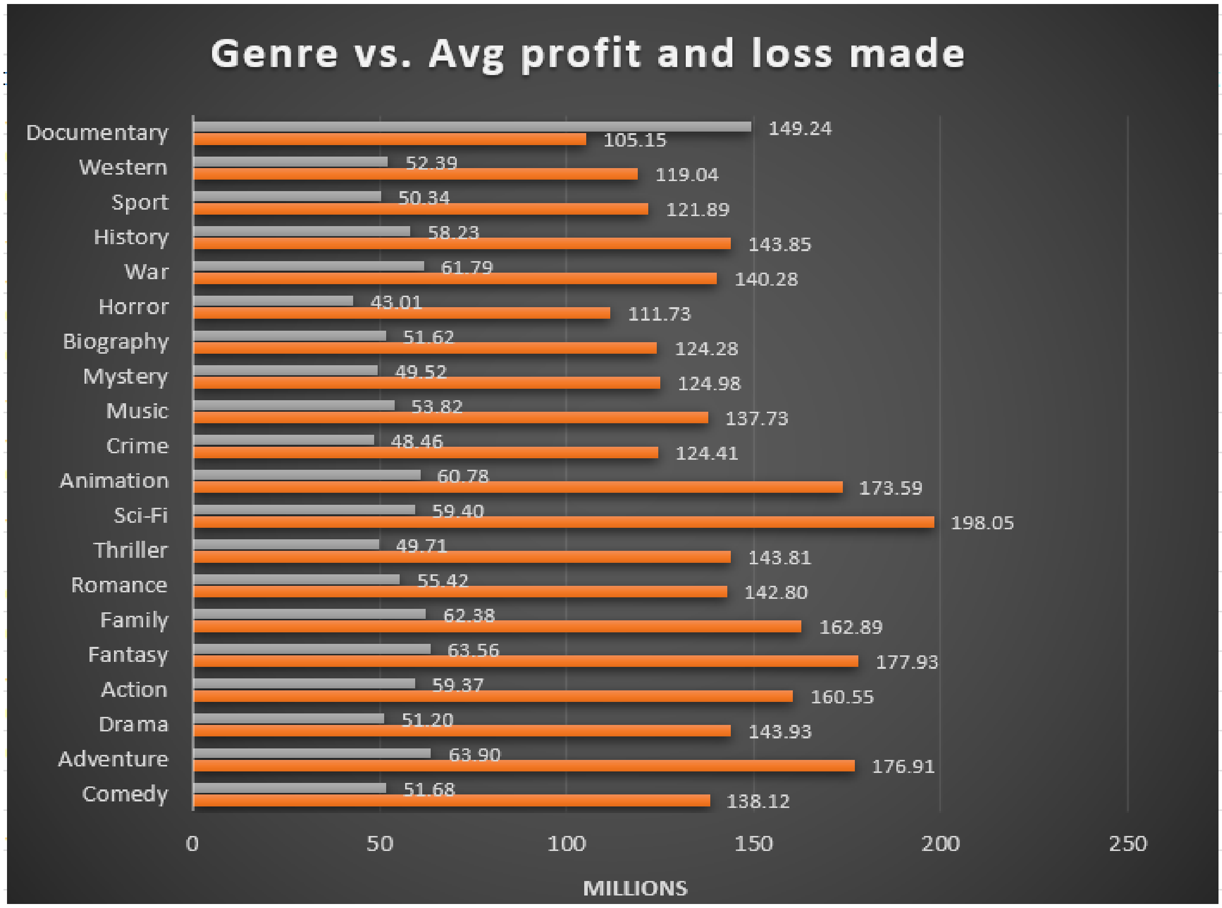
Based on the following charts, for genres, the average mean profits made seem to increase with rise in average mean budget. Genres like Sci-Fi, Action, Adventure, Fantasy and Animation seem to make big profits especially Sci-Fi movies but also require big budgets. However, according to the chart for genre count vs. IMDB score, sci-fi, animation, fantasy and adventure also have more movies that incurred losses.

Another thing to notice is that the genres Comedy, Drama, Adventure and Action have the most movies made overall and therefore share similar counts for most profitable and least profitable movies.

All in all, the top performing genres based on all three charts seem to be: **Sci-Fi, Adventure, Fantasy, Animation, Family and Action**.







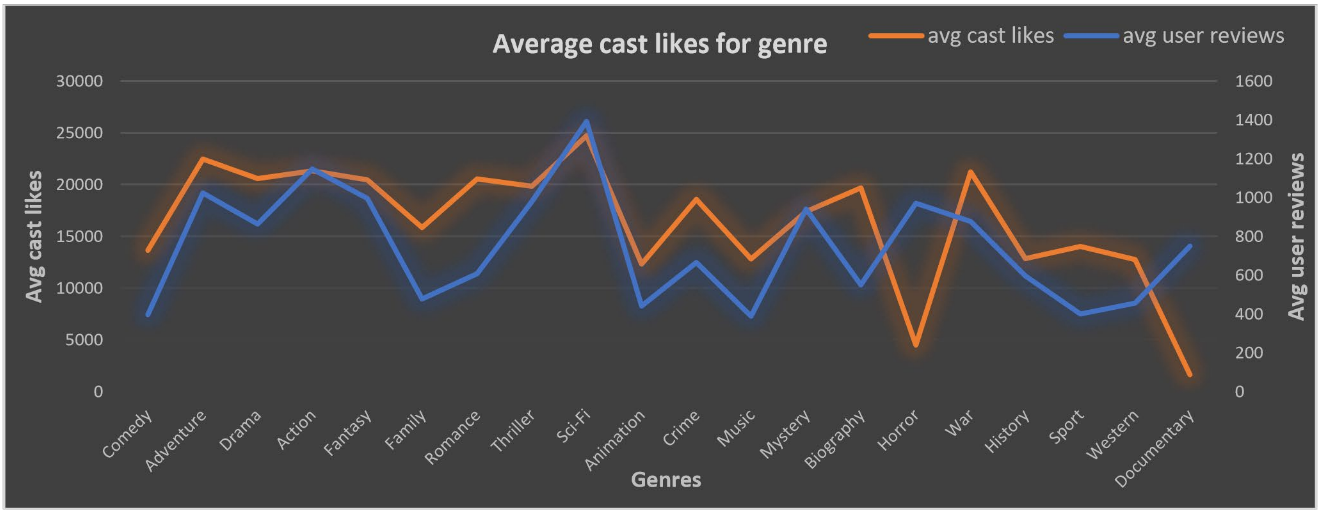
Q4. Why are genres like Sci-Fi, Action, Adventure, Fantasy, and Animation more profitable to exploit than others?

According to the list for top 20 most profitable movies, most movies belong to a combination of these genres. This may be due to a few reasons. These genres attract a vast audience as the genres come together to form a complex world engaging in fiction and very complex story-telling. They typically offer high levels of excitement, visual effects, and imaginative storytelling, which can captivate audiences and encourage repeat viewings. Movies in these genres often lead to lucrative merchandising deals, including toys, clothing, and video games, which add to their profitability. Such genres are well-suited for creating franchises and sequels, which can generate sustained revenue over time.

Movie Title	Genres
Avatar	Action-Adventure Fantasy Sci-Fi
Jurassic World	Action-Adventure Sci-Fi Thriller
Titanic	Drama Romance
Star Wars: Episode IV - A New Hope	Action-Adventure Fantasy Sci-Fi
E.T. the Extra-Terrestrial	Family Sci-Fi
The Avengers	Action-Adventure Sci-Fi
The Lion King	Adventure Animation Drama Family Music
Star Wars: Episode I - The Phantom Menace	Action Adventure Fantasy Sci-Fi
The Dark Knight	Action Crime Drama Thriller
Spirited Away	Adventure Animation Family Fantasy
The Hunger Games	Adventure Drama Sci-Fi Thriller
Deadpool	Action Adventure Comedy Romance Sci-Fi
The Hunger Games: Catching Fire	Adventure Sci-Fi Thriller
Jurassic Park	Adventure Sci-Fi Thriller
The Secret Life of Pets	Animation Comedy Family
Despicable Me 2	Animation Comedy Family Sci-Fi
American Sniper	Action Biography Drama History Thriller War
Finding Nemo	Adventure Animation Comedy Family
Shrek 2	Adventure Animation Comedy Family Fantasy Romance
The Lord of the Rings: The Return of the King	Action Adventure Drama Fantasy

Q5. Why do these genres captivate audiences and encourage repeat viewings as well as franchises?

Despite their fantastical elements, these genres often explore universal themes such as heroism, good vs. evil, and personal growth. These themes resonate deeply with viewers, making the stories memorable and impactful. These genres often feature well-developed characters with compelling arcs. Audiences become emotionally invested in these characters, which encourages repeat viewings and anticipation for sequels. These genres often foster strong fan communities. Fans engage in discussions, create fan art, and participate in conventions, which enhances their connection to the movies and encourages repeat viewings. The following chart depicts the universal appeal of such genres based on their average user reviews and cast Facebook likes.



According to this chart movies with genres Sci-Fi, Action, Adventure, Fantasy seems to get the highest number of reviews from viewers and the actors seem to have a larger fanbase leading to higher amount of likes on Facebook. The genre animation has considerably low reviews and cast likes.

Q6. Why does the Animation genre have fewer user reviews and cast likes?

For most animated movies, the cast, even though they may include famous actors for voice-acting, is not the central focus since the characters are animated and only their voices are represented by the famous actors. Additionally, most animated movies target children as their primary audience. This audience may not be familiar with reviews and may not care to leave a review or research the cast of the movie. However, they are still a good target for merchandise sales, similar to other genres that garner a fanbase.

Results:

This project was very crucial in helping me understand the complexities in the world of movie production as well as the analysis that goes into predicting what makes a successful movie in terms of IMDB scores and gross revenue. A big realization during this project was that there's not a single component determining the success of a movie but rather a combination of factors. A genre on its own does not determine the success of a movie, but with a combination of factors such as the director's vision, the budget, the cast's performance, the screenplay, and even the marketing strategy, a movie can achieve significant success.

Another important aspect uncovered was the significance of targeting niche groups of consumers. By understanding and exploiting the interests of these specific audiences, movies can achieve substantial success even without broad appeal.

Furthermore, the project highlighted the financial benefits of movie franchises. Franchises tend to bring in more money due to their established fan base, brand recognition, and the ability to create a series of interconnected stories that keep audiences engaged over multiple installments.