IMDB Movie Analysis

1. Project Description:

1.1 Overview: The project centers around analyzing factors that influence the success of movies on IMDB. The primary goal is to uncover insights into how different variables such as genre, duration, language, director, and budget impact IMDB ratings. By conducting comprehensive data analysis, the project aims to provide actionable insights for stakeholders in the movie industry.

1.2 Objectives:

1. Analyze the distribution of movie *genres* and their correlation with IMDB scores.
2. Investigate the relationship between movie *duration* and IMDB ratings.
3. Examine the influence of movie *language* on IMDB ratings.
4. Evaluate the impact of *directors* on IMDB scores based on their average ratings.
5. Explore the correlation between movie *budgets, gross earnings, and profit margins*.

2. Approach:

2.1 Data Collection and Preparation:

- Dataset: Utilized a comprehensive dataset sourced from IMDB, encompassing movie details such as genres, durations, languages, directors, budgets, actor names, actors' facebook likes, movie keywords, IMDB ratings, etc.
- Data Cleaning: Conducted rigorous data cleaning processes to handle missing values, remove duplicates, and ensure data integrity for accurate analysis.

  For removing duplicates, the excel's 'remove duplicates' option on the data menu was used.

  For handling missing values, skewness of the data was calculated using the SKEW() excel function, which stated the amount of skewness (outliers) a group of data contained. If the skewness was high in magnitude it suggested that the mean is highly influenced by the presence of outliers, so therefore it is not a good option to be filled as the missing value, and instead median should be used. When the skewness was low, mean was used.

	A	B	C	D	E	F	G	H
1	duration	imdb_score					descriptive stats	duration
2	178	7.9			missing		mean	107.19
3	169	7.1					median	103.00
4	148	6.8					mode	90.00
5	164	8.5					min	7.00
6	103	7.1					max	511.00
7	132	6.6					varience	633.07
8	156	6.2					std dev	25.16
9	100	7.8						
10	141	7.5						
11	153	7.5						
12	183	6.9						
13	169	6.1						
14	106	6.7						
15	151	7.3						
16	150	6.5						

Here, due to high skewness, the median was deployed to fill in the missing values for movie duration.

	E	F	G	H	I	J	K	L
				language	mean imdb	median imdb	std dev imdb	count
		missing		English	6.4	0.0	1.1	4595
				Japanese	7.3	7.5	1.0	17
				French	7.0	7.2	0.7	73
				Mandarin	6.8	7.1	1.0	24
				Aboriginal	7.0	7.0	0.8	2
				Spanish	6.9	7.2	0.9	40
				Filipino	6.7	6.7	0.0	1
				Hindi	6.6	7.0	1.4	28

Here, in language analysis, due to the count of the English language being disproportionately larger than others, therefore being the mode, it was used to fill in the missing values.

4817	Michael Mann	7.5
4818	Joe Carnahan	7.6
4819	Ridley Scott	8.6
4820	Mark Steven	8.8
4821	Nora Ephron	7.6
4822	Francis Lawrence	7.5
4823	Daniel Espinosa	8.3
4824	Tim Burton	7.5
4825	Walter Hill	7.4
4826	Ron Howard	8.1
4827	William Friedkin	7.3
4828	Tibor Takács	6.6
4829	Michael Patrick	7
4830	Peter Segal	6.7
4831	Jonathan Glassner	8.4
4832	Andy Wilson	7.9
4833	Roger Bamford	7.8
4834	Philip Kaufman	7.5
4835	Edward Zwick	7
4836	Richard Laxton	8.5
4837	Fabrice Gobert	8.3
4838	Bryan Spicer	7.5
4839	Steve Gordon	7.4
4840	Terry Hughes	7.8
4841	Brett Ratner	5.8
4842	James LaRosa	7
4843	Brian Kirk	8.6
4844	Peter Berg	8.7
4845	Luc Besson	7.5

Here, for the director analysis, the missing values of the directors was filled using domain knowledge and AI assistance to ensure accurate replacements, enhancing dataset integrity and completeness

2.2 Analytical Techniques:

A. **Movie Genre Analysis**: Employed Excel functions (COUNTIF, AVERAGE, MEDIAN, MODE, etc.) to determine the prevalence of genres and their impact on IMDB scores.

	genre1	genre2	genre3	genre4	genre5	genre6	genre7	imdb_score
2	Action	Adventure	Fantasy	Sci-Fi				7.9
3	Action	Adventure	Fantasy					7.1
4	Action	Adventure	Thriller					6.8
5	Action	Thriller						8.5
6	Documentary							7.1
7	Action	Adventure	Sci-Fi					6.6
8	Action	Adventure	Romance					6.2
9	Adventure	Animation	Comedy	Family	Fantasy	Musical	Romance	7.8
10	Action	Adventure	Sci-Fi					7.5
11	Adventure	Family	Fantasy	Mystery				7.5
12	Action	Adventure	Sci-Fi					6.9
13	Action	Adventure	Sci-Fi					6.1
14	Action	Adventure						6.7
15	Action	Adventure	Fantasy					7.3
16	Action	Adventure	Western					6.5
17	Action	Adventure	Fantasy	Sci-Fi				7.2
18	Action	Adventure	Family	Fantasy				6.6
19	Action	Adventure	Sci-Fi					8.1
20	Action	Adventure	Fantasy					6.7
21	Action	Adventure	Comedy	Family	Fantasy	Sci-Fi		6.8
22	Adventure	Fantasy						7.5
23	Action	Adventure	Fantasy					7
24	Action	Adventure	Drama	History				6.7
25	Adventure	Fantasy						7.9
26	Adventure	Family	Fantasy					6.1
27	Action	Adventure	Drama	Romance				7.2
28	Drama	Romance						7.7
29	Action	Adventure	Sci-Fi					8.2
30	Action	Adventure	Sci-Fi	Thriller				5.9

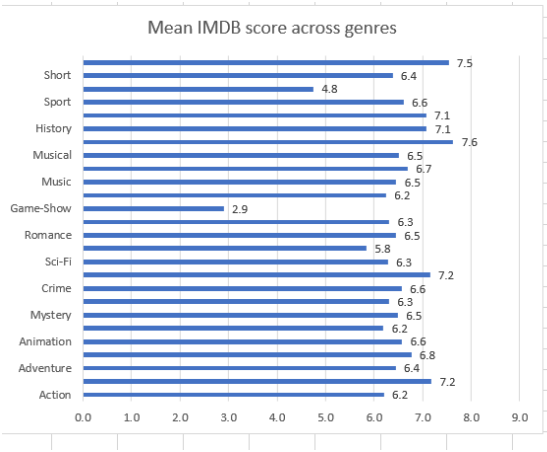
unique genres	count
Action	1153
Documentary	121
Adventure	923
Drama	2594
Animation	242
Comedy	1872
Mystery	500
Fantasy	610
Crime	889
Biography	293
Sci-Fi	616
Horror	565
Romance	1106
Thriller	1408
Game-Show	1
Family	546
Music	214
Western	97
Musical	132
Film-Noir	6
History	207
War	213
Sport	182
Reality-TV	2
Short	5
News	3

most common genres	count
Drama	2594
Comedy	1872
Thriller	1408
Action	1153
Romance	1106

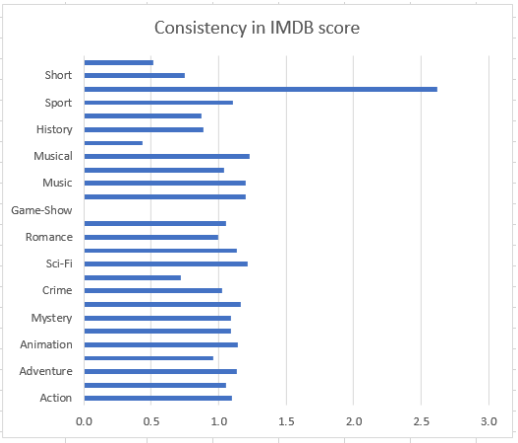
First, I split the genres into different columns, determined the unique genres using the 'delete duplicates' option, and calculated the counts for every unique genre. Then I used the LARGE and MAX functions to calculate the top 5 most common genres.

unique genre	mean imdb	median imdb	mode imdb	min imdb	max imdb	varience imdb	std dev imdb	count
Action	6.2	6.3	6.1	1.7	9.1	1.3	1.1	1153
Documentary	7.2	7.4	7.5	1.6	8.7	1.1	1.1	121
Adventure	6.4	6.6	6.7	1.9	8.9	1.3	1.1	923

I then calculated the descriptive statistics for every genre, and visualised different relationships.

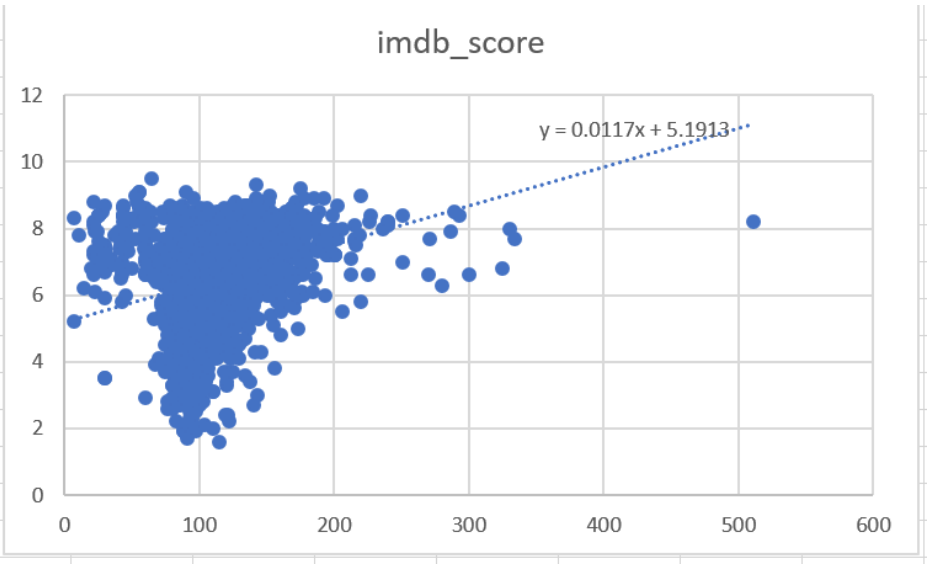


unique genre	std dev imdb
Action	1.1
Documentary	1.1
Adventure	1.1
Drama	1.0
Animation	1.1
Comedy	1.1
Mystery	1.1
Fantasy	1.2
Crime	1.0
Biography	0.7
Sci-Fi	1.2
Horror	1.1
Romance	1.0
Thriller	1.1
Game-Show	0.0
Family	1.2
Music	1.2
Western	1.0
Musical	1.2
Film-Noir	0.4
History	0.9



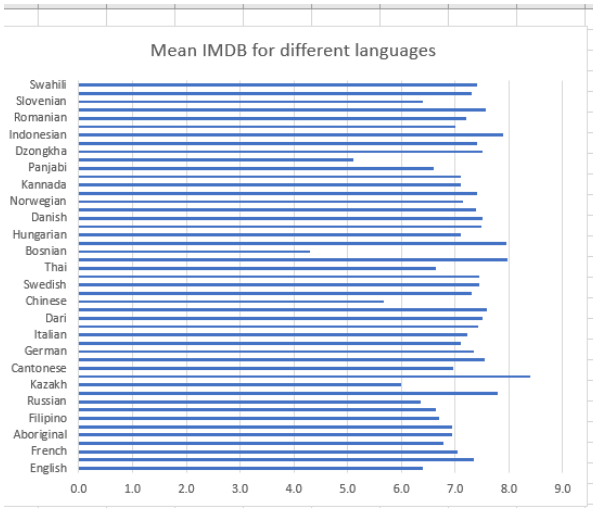
B. Movie Duration Analysis: Calculated descriptive statistics (mean, median, standard deviation) and employed scatter plots with trendlines to explore the relationship between movie duration and IMDB ratings.

descriptive stats	duration
mean	107.19
median	103.00
mode	90.00
min	7.00
max	511.00
varience	633.07
std dev	25.16

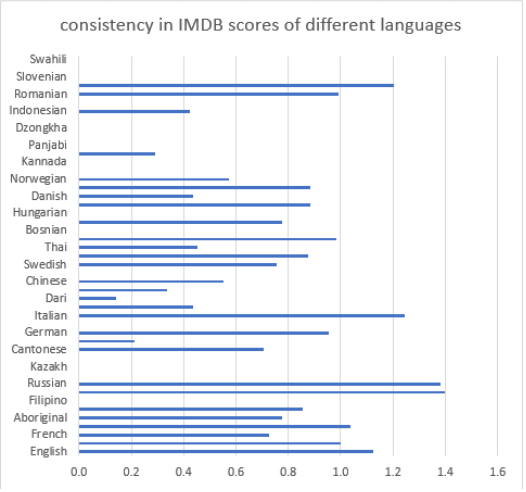


C. Language Analysis: Used Excel functions (COUNTIF, AVERAGE, STDEV) to calculate the descriptive statistics of every unique language and explored different relationships using visualizations to assess the influence of languages on IMDB scores.

language	mean imdb	median imdb	std dev imdb	count
English	6.4	0.0	1.1	4595
Japanese	7.3	7.5	1.0	17
French	7.0	7.2	0.7	73
Mandarin	6.8	7.1	1.0	24
Aboriginal	7.0	7.0	0.8	2
Spanish	6.9	7.2	0.9	40
Filipino	6.7	6.7	0.0	1
Hindi	6.6	7.0	1.4	28
Russian	6.4	6.5	1.4	11
Maya	7.8	7.8	0.0	1
Kazakh	6.0	6.0	0.0	1
Telugu	8.4	8.4	0.0	1
Cantonese	7.0	7.2	0.7	11
Icelandic	7.6	8.1	0.2	2
German	7.3	7.6	1.0	19
Aramaic	7.1	7.1	0.0	1



language	std dev imdb
English	1.1
Japanese	1.0
French	0.7
Mandarin	1.0
Aboriginal	0.8
Spanish	0.9
Filipino	0.0
Hindi	1.4
Russian	1.4
Maya	0.0
Kazakh	0.0
Telugu	0.0
Cantonese	0.7
Icelandic	0.2
German	1.0
Aramaic	0.0
Italian	1.2
Dutch	0.4
Dari	0.1
Hebrew	0.3
Chinese	0.6
Mongolian	0.0
Swedish	0.8

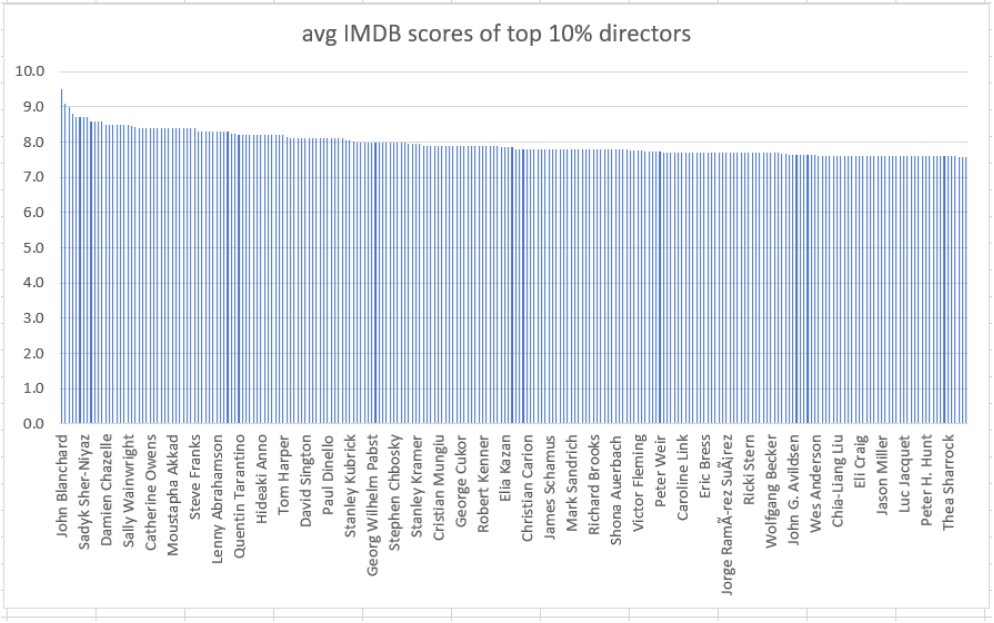


D. Director Analysis: Applied Excel's AVERAGE and PERCENTILE functions to identify top directors based on IMDB ratings and analyze their contribution to movie success.

director_name	mean imdb	count		90th %ile	Top or Below 10%	Top 10% directors	mean imdb
John Blanchard	9.5	1.0		7.5675	Top 10%	John Blanchard	9.5
Krzysztof Kieślowski	9.1	1.0			Top 10%	Krzysztof Kieślowski	9.1
Noah Hawley	9.0	1.0			Top 10%	Noah Hawley	9.0
Rob McElhenney	8.8	1.0			Top 10%	Rob McElhenney	8.8
Cary Bell	8.7	1.0			Top 10%	Cary Bell	8.7
Mitchell Altieri	8.7	1.0			Top 10%	Mitchell Altieri	8.7
Sadyk Sher-Niyaz	8.7	1.0			Top 10%	Sadyk Sher-Niyaz	8.7
Stefano Sollima	8.7	1.0			Top 10%	Stefano Sollima	8.7
Brian Kirk	8.6	1.0			Top 10%	Brian Kirk	8.6
Charles Chaplin	8.6	1.0			Top 10%	Charles Chaplin	8.6

I first calculated the mean IMDB scores for every unique director using pivot table and chart. Then I calculated the 90th percentile using the PERCENTILE function which came out to be 7.5675, then listed the directors which came out to be in the Top 10%.

Then, I compared the distribution of scores from the Top 10% directors with the whole distribution.



AB	AC	AD	AE
average IMDB of all directors		std dev of all directors	
6.3		1.1	
average IMDB of top 10% directors		std dev of top 10% directors	
7.9		0.3	

E. Budget Analysis: Utilized Excel's CORREL function to determine the correlation between movie budgets and gross earnings, calculated the profit margins and then identified movies with the highest profit margins using the MAX function.

			profit margin	top 10 highest profit margins	movie name
			523505847	523505847	AvatarÂ
9800000	missing budget values		9404152	502177271	Jurassic WorldÂ
5043962	missing gross values		-44925825	458672302	TitanicÂ
			198130642	449935665	Star Wars: Episode IV - A New HopeÂ
	correlation coefficient:	0.239578	5243962	424449459	E.T. the Extra-TerrestrialÂ
			-190641321	403279547	The AvengersÂ
			78530303	377783777	The Lion KingÂ
			-59192738	359544677	Star Wars: Episode I - The Phantom MenaceÂ
			208991599	348316061	The Dark KnightÂ
			51956980	329999255	The Hunger GamesÂ

3. Tech-Stack Used

Software:

- Microsoft Excel 2022:

Microsoft Excel 2022 was the primary tool used for data analysis in this project. Its robust capabilities for data manipulation, statistical analysis, and visualization were essential in managing and interpreting the large dataset. Excel's functions were used extensively to calculate descriptive statistics, correlation coefficients, percentiles, and to create visualizations such as scatter plots and trendlines. The software's filtering and conditional formatting features facilitated the identification and handling of missing values, ensuring the integrity and completeness of the dataset.

AI Tool:

- ChatGPT by OpenAI:

ChatGPT was used to assist in filling missing director values by providing accurate and contextually relevant replacements. This tool helped ensure the completeness of the dataset and maintained the integrity of the data analysis process.

4. Insights

4.1 Insights and Knowledge Gained:

Through the comprehensive analysis of the IMDB movie dataset, several key insights and trends were discovered, providing a deeper understanding of the factors influencing movie success on IMDB.

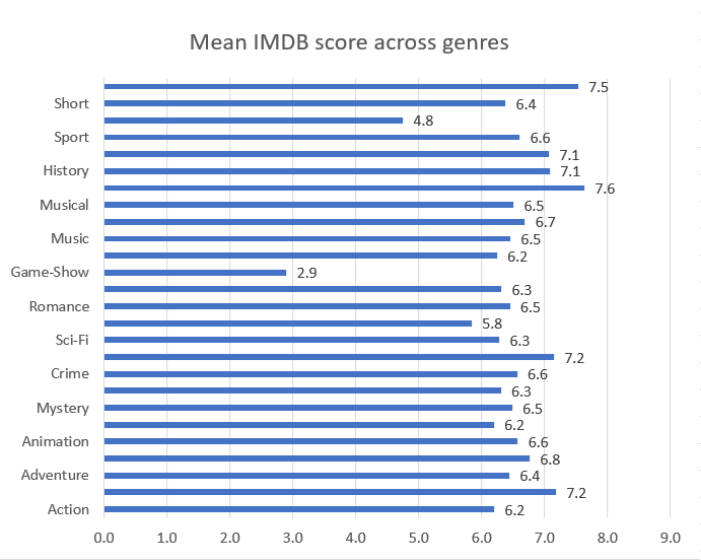
A. Movie Genre Analysis

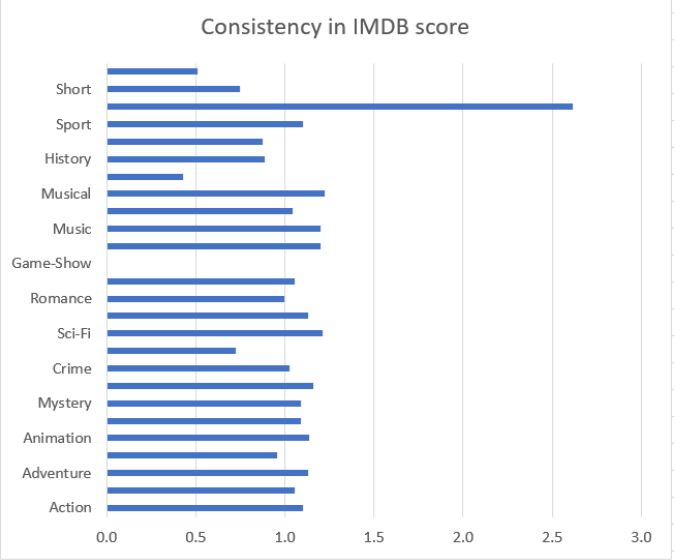
- Most Common Genres: The most prevalent genres in the dataset were Drama, Comedy, Thriller, Action, and Romance.

- Impact on IMDB Scores:

- Highest Mean IMDB Scores: Genres such as Film-Noir, News, Biography, and Documentary had the highest mean IMDB scores, indicating that these genres are generally well-received by audiences.

- Consistency in Scores: Film-Noir, News, and Biography also had the lowest standard deviations in their IMDB scores, suggesting consistent audience appreciation for these genres.





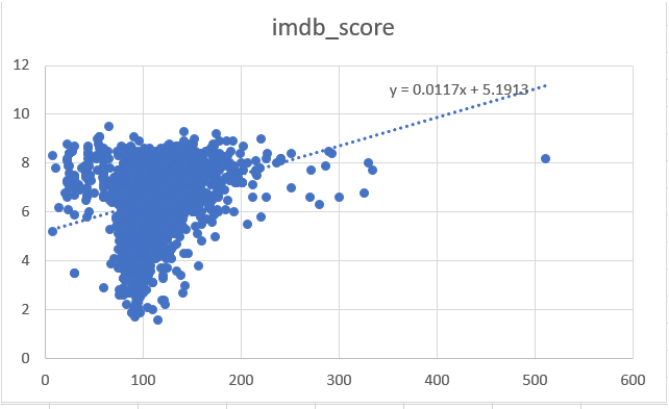
B. Movie Duration Analysis

- Distribution of Durations: The analysis revealed the overall distribution of movie durations and their descriptive statistics.

- Relationship with IMDB Scores:

- Positive Correlation: A positive slope in the trendline of the scatter plot suggested that longer movies tend to have higher IMDB scores.

- Regression Equation: The relationship can be expressed as: IMDB Score = (0.0117) Duration + 5.1913, indicating a positive but modest impact of duration on ratings.



C. Language Analysis

- Most Common Languages: English dominated the dataset with 4595 movies, significantly outnumbering other languages.

- Impact on IMDB Scores:

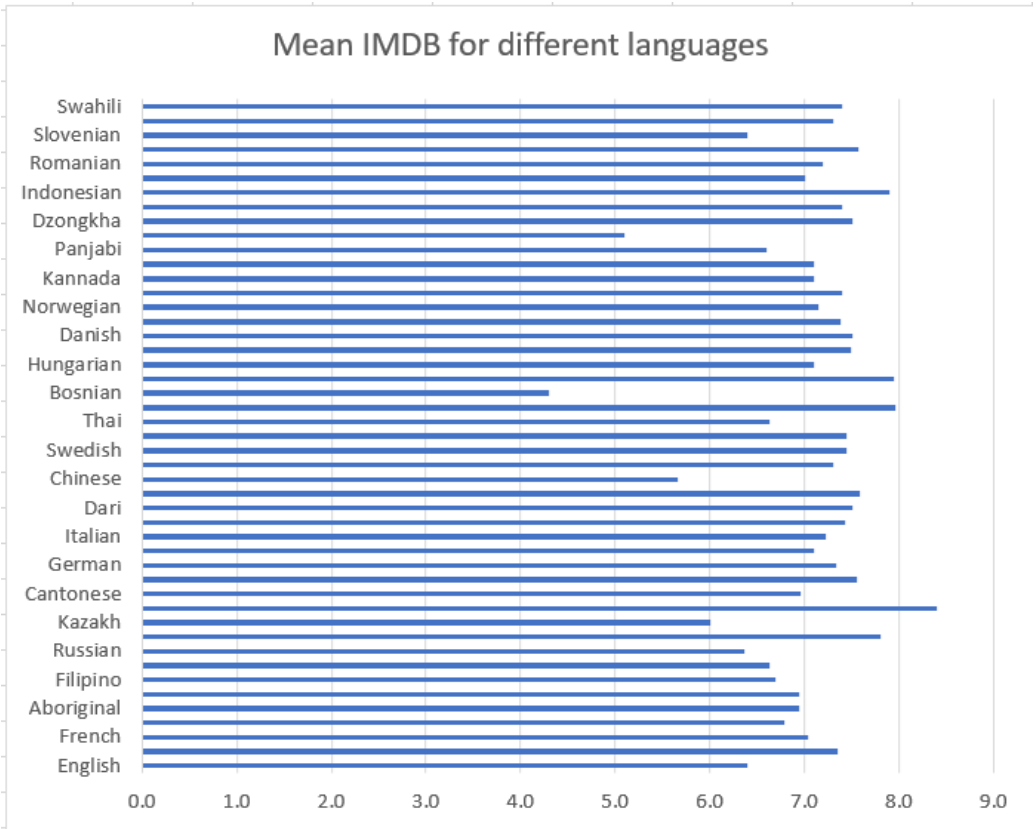
- Highest Average IMDB Scores: Telegu, Polish, and Indonesian were the top three languages with the highest average IMDB scores.

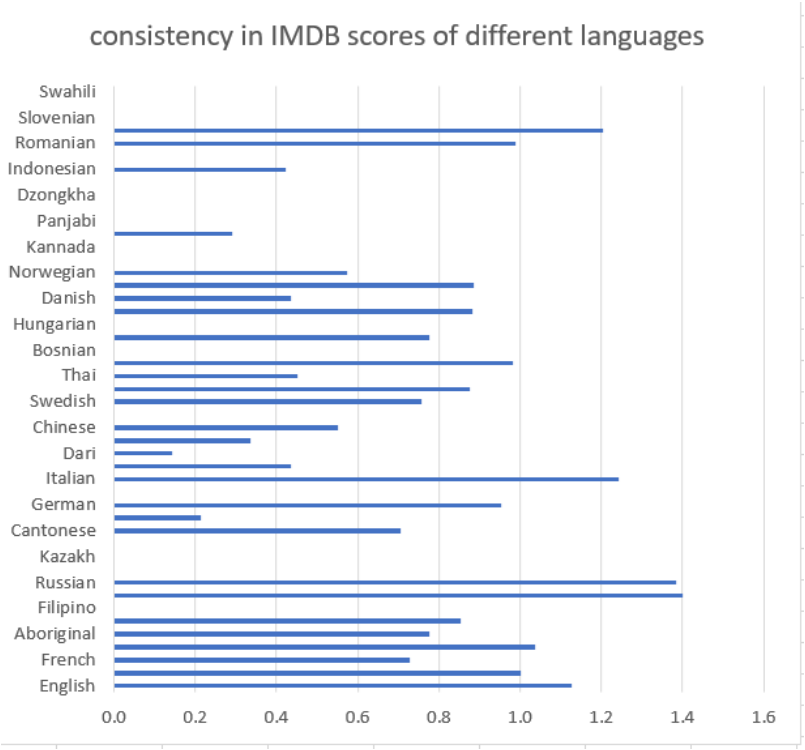
- Lowest Average IMDB Scores: Bosnian, Tamil, and Chinese had the lowest average IMDB scores.

- Consistency and Variability:

- Most Variable: Russian, Hindi, and Italian had the highest variability in scores, indicating inconsistent reception.

- Least Variable: Dari, Icelandic, Zulu, and Hebrew had the least variability, suggesting consistent audience ratings.





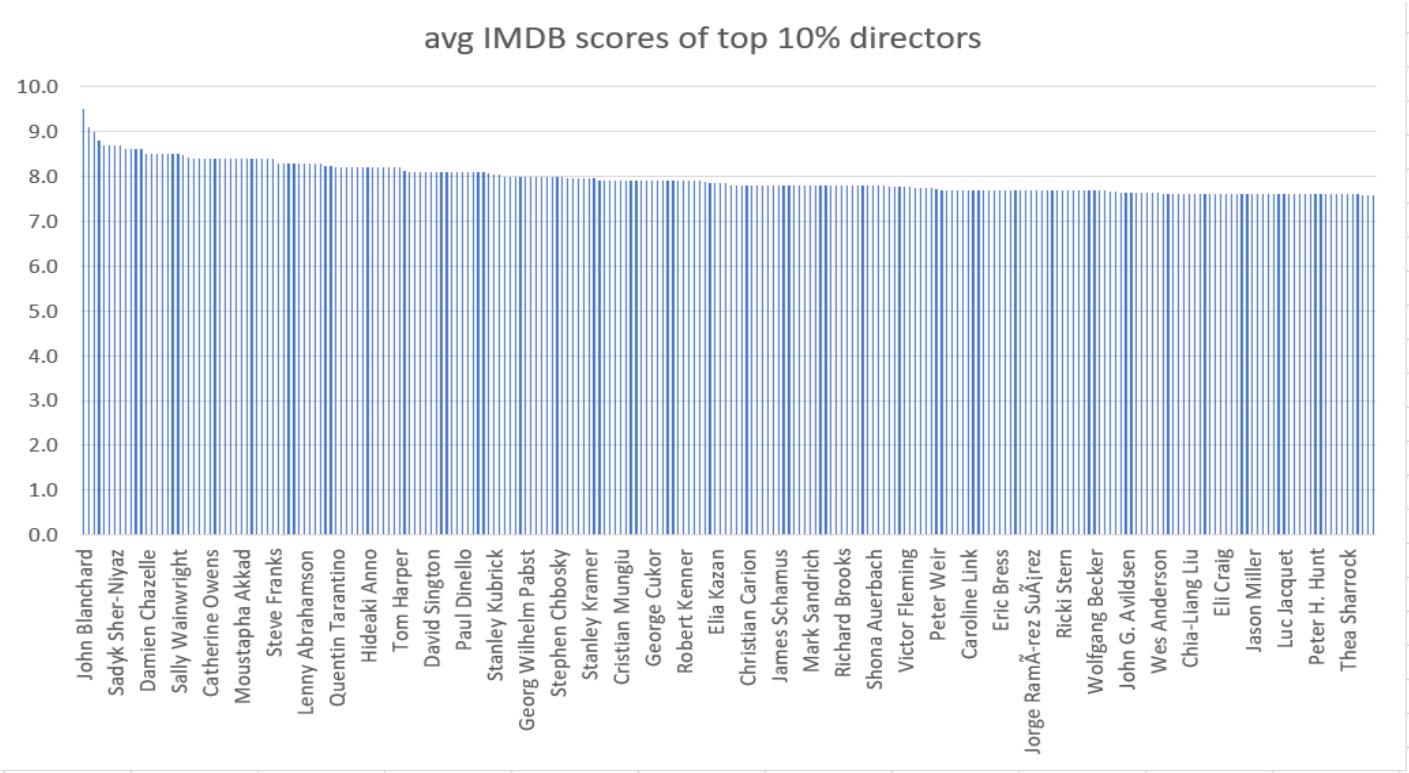
D. Director Analysis

- Overall vs. Top 10% Directors:

- Average IMDB Scores: The average IMDB score for all directors was 6.3, whereas the top 10% of directors had an average score of 7.9.

- Standard Deviation: The standard deviation for all directors was 1.1, while for the top 10%, it was 0.3, indicating that top directors consistently received higher ratings.

- Top Directors: Percentile analysis highlighted directors who significantly contributed to the success of movies, showing a clear difference in ratings distribution.



E. Budget Analysis

- Correlation Analysis:

- Correlation Coefficient: The correlation coefficient between movie budgets and gross earnings was calculated as 0.2396, suggesting a positive but weak correlation.

- Profit Margin Analysis:

- Top 10 Movies with Highest Profit Margins: The movies with the highest profit margins were:

  - Avatar

  - Jurassic World

  - Titanic

  - Star Wars: Episode IV - A New Hope

  - E.T. the Extra-Terrestrial

  - The Avengers

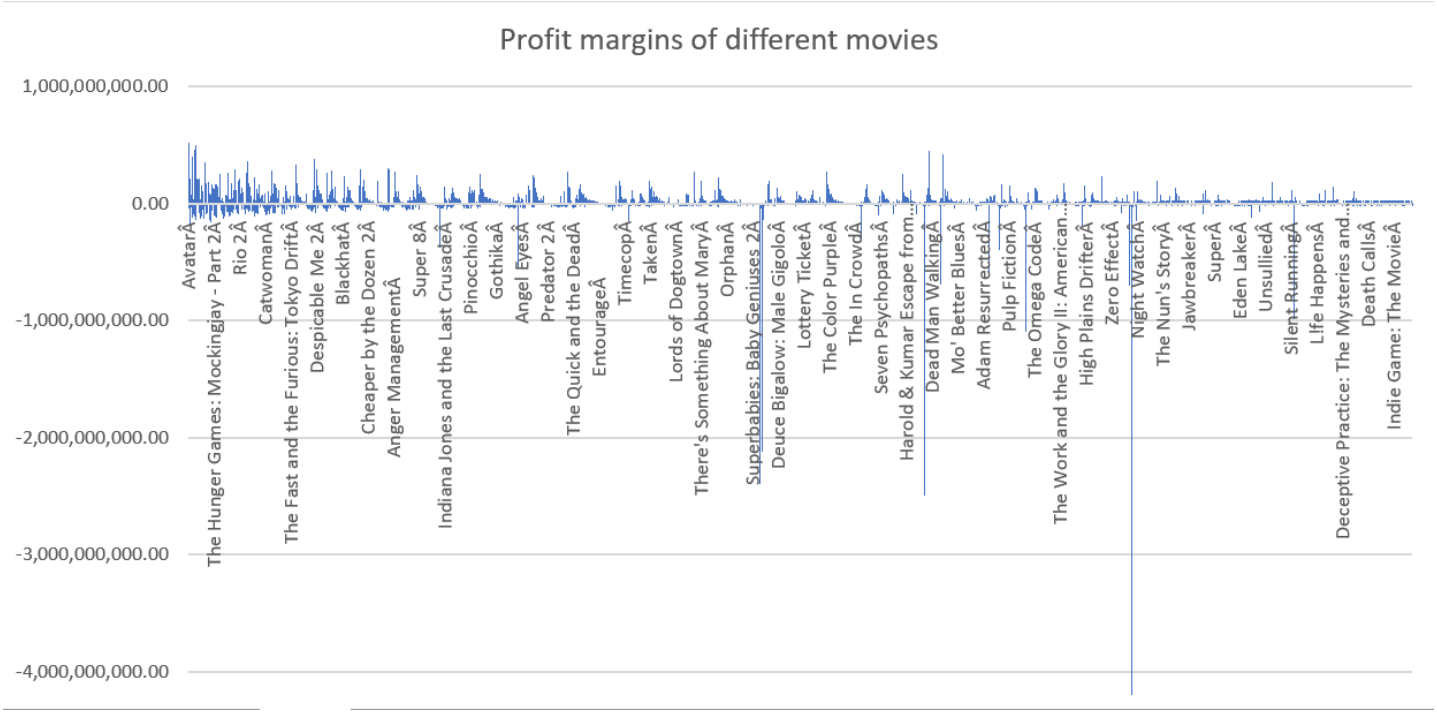
  - The Lion King

  - Star Wars: Episode I - The Phantom Menace

  - The Dark Knight

  - The Hunger Games

These movies demonstrated significant financial success, reflecting the importance of budget management and audience appeal in achieving high profit margins.



4.2 Key Findings and Trends

Storytelling and Genre:

- Complex and engaging storytelling, often found in drama films, resonates well with audiences, leading to higher ratings. The consistency of high ratings in genres such as Film-Noir, News, and Biography indicates that these genres, despite being less common, maintain a dedicated and appreciative audience base.

- *Why do Film-Noir, News, and Biography genres have the highest mean IMDB scores?*

  These genres often feature complex, thought-provoking content that appeals to niche audiences.

- *Why does complex, thought-provoking content appeal to niche audiences?*

  Niche audiences tend to appreciate depth and intellectual stimulation in films.

- *Why do niche audiences appreciate depth and intellectual stimulation?*

  These audiences seek meaningful and enriching experiences from movies, beyond mere entertainment.

- *Why do meaningful and enriching experiences lead to higher IMDB scores?*

  Viewers are likely to give higher ratings to movies that provide a deeper connection and intellectual satisfaction.

- *Why do higher ratings from niche audiences matter?*

  High ratings from dedicated viewers boost the movie's reputation, attracting more viewers who trust these reviews.

Optimal Movie Duration:

- There is an optimal range for movie duration that maximizes audience engagement and ratings. The positive correlation between movie duration and IMDB scores suggests that longer movies, which can develop more detailed plots and characters, tend to receive higher ratings.

- *Why do longer movies tend to have higher IMDB scores?*

  Longer movies have more time to develop plots and characters fully.

- *Why does fully developing plots and characters lead to higher IMDB scores?*

  Well-developed plots and characters create a richer and more immersive storytelling experience.

- *Why does a richer storytelling experience result in higher ratings?*

  Viewers feel more engaged and emotionally connected, leading to a more satisfying viewing experience.

- *Why does a satisfying viewing experience translate to higher IMDB scores?*

Satisfied viewers are more likely to leave positive reviews and recommend the movie to others.

- *Why are positive reviews and recommendations important for IMDB scores?*

  Positive reviews and recommendations increase the film's visibility and credibility, drawing in more viewers and improving its overall ratings.

Language and Accessibility:

- English-language films dominate the dataset and generally perform better due to their broader accessibility and larger target audience. However, high-quality non-English films, particularly in languages like Telegu, Polish, and Indonesian, also achieve significant success, demonstrating that well-crafted stories can transcend language barriers.

- *Why do Telegu, Polish, and Indonesian films have high average IMDB scores?*

  These films may offer unique cultural perspectives and high-quality storytelling that resonate with viewers.

- *Why do unique cultural perspectives and high-quality storytelling resonate with viewers?*

  Viewers appreciate authenticity and originality, which stand out in a diverse film landscape.

- *Why does authenticity and originality in films lead to higher IMDB scores?*

  Authentic and original films provide fresh and memorable experiences for the audience.

- *Why do fresh and memorable experiences result in higher ratings?*

  Viewers are likely to give higher scores to films that leave a lasting impact and provide a unique viewing experience.

- *Why are higher scores important for films in less common languages?*

  High scores increase the visibility and recognition of these films, attracting a broader audience and promoting cultural diversity in cinema.

Directorial Impact:

- Directors play a crucial role in determining a movie's success. The analysis showed that top directors consistently produce highly rated films, emphasizing the importance of directorial vision and expertise in filmmaking. Directors in the top percentile of ratings significantly outperform the average, highlighting their critical contribution to a movie's quality and reception.

- *Why do movies directed by top-tier directors have higher average IMDB scores?*

  Top-tier directors have extensive experience and a strong vision for their projects.

- *Why does having extensive experience and a strong vision lead to higher IMDB scores?*

  Experienced directors can create more engaging and cohesive stories that resonate well with audiences.

- *Why do engaging and cohesive stories resonate well with audiences?*

  Engaging stories capture the audience's attention and create emotional connections, leading to a more enjoyable viewing experience.

- *Why does a more enjoyable viewing experience result in higher IMDB scores?*

  When viewers have a positive experience, they are more likely to rate the movie highly and recommend it to others.

- *Why are positive ratings and recommendations important for a movie's success?*

  Positive ratings and recommendations increase the movie's visibility and credibility, attracting more viewers and enhancing its overall success.

Strategic Budgeting:

- Effective budget management is critical for financial success. The correlation between movie budgets and gross earnings, although modest, indicates that higher budgets can lead to higher earnings. Additionally, movies like Avatar and Jurassic World have shown that substantial profit margins can be achieved through efficient resource utilization, highlighting the importance of strategic budgeting in the film industry.

- *Why do movies with higher budgets tend to have higher ratings?*

  Higher budgets allow for better production quality, including special effects, sets, and talent.

- *Why does better production quality lead to higher ratings?*

  High production quality enhances the viewer's experience, making the movie more visually appealing and immersive.

- *Why does an enhanced viewer experience lead to higher ratings?*

  Viewers are more likely to rate a movie highly if they enjoyed watching it and were impressed by its production values.

- *Why are viewers more likely to rate a movie highly if they enjoyed watching it?*

  Positive experiences lead to positive reviews, and viewers often share their satisfaction through high ratings.

- *Why do positive reviews matter?*

  Positive reviews influence other viewers' decisions to watch the movie, increasing its popularity and success. Additionally, they contribute to a higher overall IMDB rating for the film.

4.3 Summary

Our analysis reveals actionable insights across various facets of movie production and performance. Genre analysis highlights that genres such as Film-Noir, News, Biography, and Documentary consistently yield higher IMDB scores, suggesting a strategic focus for producers aiming to maximize audience reception.

Regarding movie duration, our findings indicate that longer films tend to garner higher ratings, signaling an opportunity for filmmakers to explore extended storytelling for enhanced audience engagement.

Language diversity analysis underscores the potential of non-English films, particularly in Telegu, Polish, and Indonesian, which achieve commendable IMDB ratings, presenting an avenue for expanding audience reach and impact.

Directorial impact analysis emphasizes the pivotal role of experienced directors in creating successful movies, urging stakeholders to prioritize collaborations with top-tier talent.

Furthermore, our budget analysis highlights that strategic budgeting can lead to both higher ratings and substantial profit margins, advising prudent allocation of resources to optimize production quality and financial returns.

5. Result

Through this project, I have achieved a comprehensive understanding of the factors influencing movie success on IMDB. By analyzing genres, movie durations, languages, directorial impacts, and budgeting strategies, I've uncovered actionable insights that can inform decision-making in the film industry. My findings highlight the significance of genre selection, optimal movie durations, and the impact of directorial expertise on audience reception and ratings. Additionally, my exploration into budgeting strategies underscores the critical link between financial management and both artistic and financial success in filmmaking. This project has not only enhanced my proficiency in data analysis using tools like Excel but has also deepened my insight into the complex dynamics shaping movie ratings and profitability.

Link to the excel file:

https://docs.google.com/spreadsheets/d/19QQBhO1wOrz3_gNnfd7kLAfijAvUgfUF/edit?usp=sharing&ouid=113270886502859747924&rtpof=true&sd=true