IMDB Movie Analysis

Project Description

With this project my aim is to explore the factors influencing a movie's success on IMDB, focusing on high IMDB ratings as a measure of success. This analysis is vital for stakeholders like producers, directors, and investors seeking to understand the determinants of a movie's success for informed decision-making in future projects. By analyzing a dataset containing movie attributes such as genre, budget, duration, cast, director, and IMDB ratings, I'll uncover patterns, correlations, and key predictors of high IMDB ratings. Insights gained will aid stakeholders in optimizing their strategies for producing successful movies and maximizing returns on investment.

Approach

In initiating the analysis, I commenced by thoroughly examining the dataset's features and their respective statistics. Upon this initial exploration, I discerned certain columns that appeared to be less relevant to the investigation at hand. Consequently, I made the decision to drop the following columns: "Color," "Face No. in poster," "movie IMDB link," "plot keywords," "Director Facebook likes," "Actor 1 Facebook likes," "Actor 2 Facebook likes," and "Actor 3 Facebook likes."

Following this data refinement step, I proceeded to scrutinize the dataset for any instances of missing or null values. Subsequently, I executed the removal of these null values to ensure data integrity and reliability.

One peculiar observation I made during this phase was the presence of an extraneous character "Â" at the end of each movie name. To rectify this inconsistency, I employed the replace function to systematically eliminate this character across all relevant entries.

With these preliminary data preprocessing steps completed, I was poised to delve into addressing the core problem statement and extracting meaningful insights from the dataset.

Tech Stack used

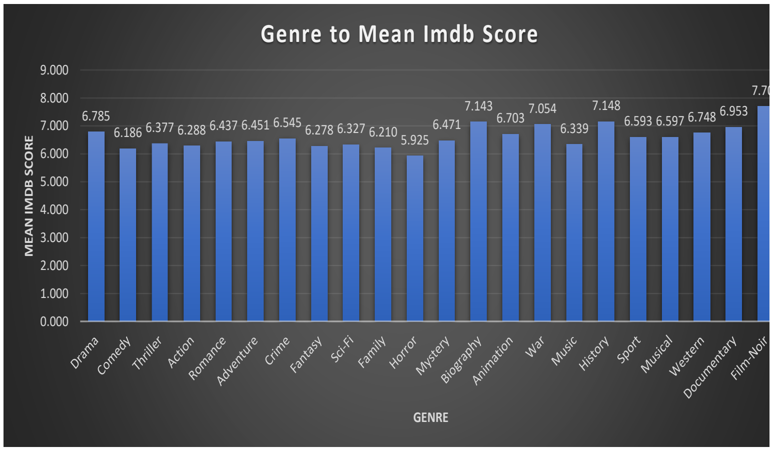
I have used Excel version 2021 to cater to this project because it has powerful tools like Pivot table and appealing chart types for visualization.

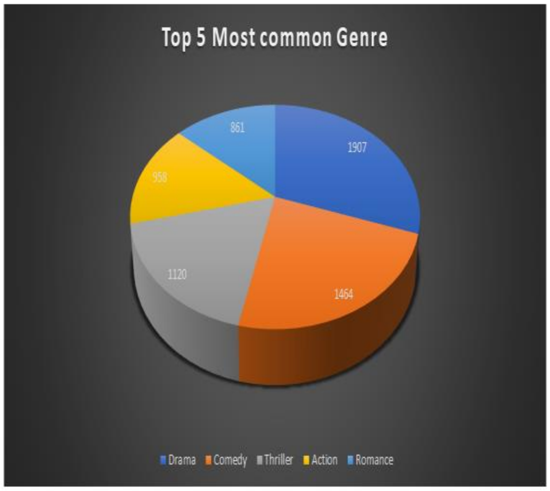
Q No. 1

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. Hint: Use Excel's COUNTIF function to count the number of movies for each genre. You might need to manipulate the 'genres' column to separate multiple genres for a single movie. Use Excel's functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV to calculate descriptive statistics. Compare the statistics to understand the impact of genre on movie ratings.

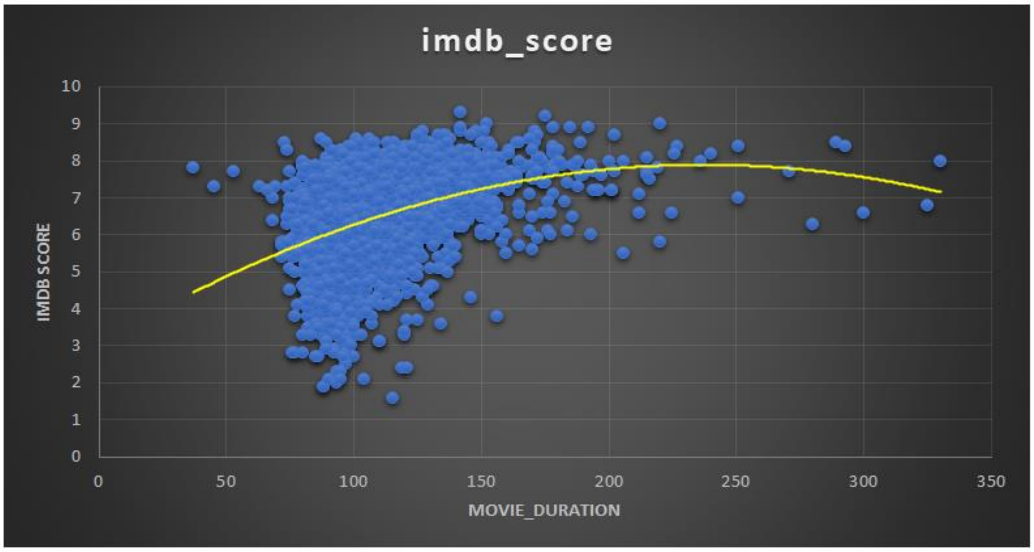
Genre	Count	Mean	Median	Mode	Max	Min	Var	Stdev
Drama	1907	6.78495	6.9	6.7	9.3	2.1	0.80886572	0.89936963
Comedy	1464	6.186066	6.3	6.7	8.8	1.9	1.07394104	1.03631126
Thriller	1120	6.376607	6.4	6.5	9	2.7	0.94392056	0.97155574
Action	958	6.288137	6.3	6.6	9	2.1	1.07962996	1.03905244
Romance	861	6.436702	6.5	6.5	8.5	2.1	0.9107212	0.95431714
Adventure	779	6.451469	6.6	6.6	8.9	2.3	1.24053173	1.1137916
Crime	712	6.544522	6.6	6.6	9.3	2.4	0.96964647	0.98470628
Fantasy	506	6.277603	6.4	6.7	8.9	2.2	1.29075724	1.13611497
Sci-Fi	494	6.326962	6.4	6.7	8.8	1.9	1.34527966	1.15986191
Family	441	6.210158	6.3	5.4	8.6	1.9	1.35218165	1.16283346
Horror	393	5.925191	6	5.9	8.6	2.3	0.99449135	0.99724187
Mystery	384	6.470649	6.5	6.6	8.6	3.1	1.033954	1.01683529
Biography	242	7.142562	7.2	7	8.9	4.5	0.50419756	0.7100687
Animation	195	6.702551	6.8	6.7	8.6	2.8	0.97912166	0.98950577
War	154	7.053896	7.1	7.1	8.6	4.3	0.63884093	0.79927525
Music	152	6.339474	6.5	6.5	8.5	1.6	1.49565005	1.22296772
History	150	7.148	7.2	7.7	8.9	5.5	0.45620403	0.67542877
Sport	148	6.593243	6.8	7.2	8.3	2	1.0851241	1.0416929
Musical	95	6.596875	6.75	7.1	8.5	2.1	1.21420066	1.10190774
Western	59	6.748333	6.75	6	8.9	4.1	0.97406497	0.9869473
Document	47	6.953191	7.4	6.6	8.5	1.6	1.91123959	1.38247589
Film-Noir	1	7.7	7.7	#N/A	7.7	7.7	#DIV/0!	#DIV/0!

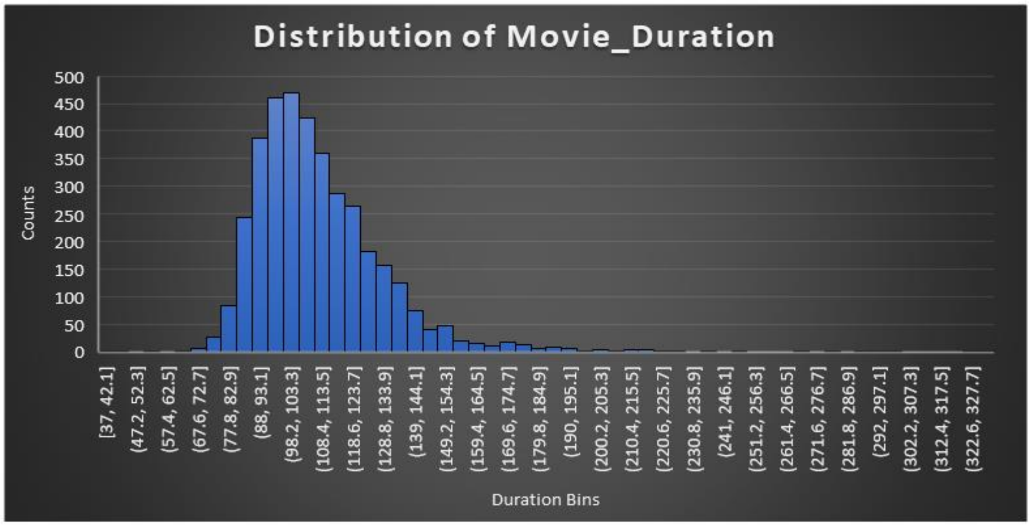




Q No. 2	B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score. Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score. Hint: Calculate descriptive statistics such as mean, median, and standard deviation for movie durations. Use Excel's functions like AVERAGE, MEDIAN, and STDEV. Create a scatter plot to visualize the relationship between movie duration and IMDB score. Add a trendline to assess the direction and strength of the relationship.

Stats	Value
Mean	110.2282
Median	106
Stdev	22.64431

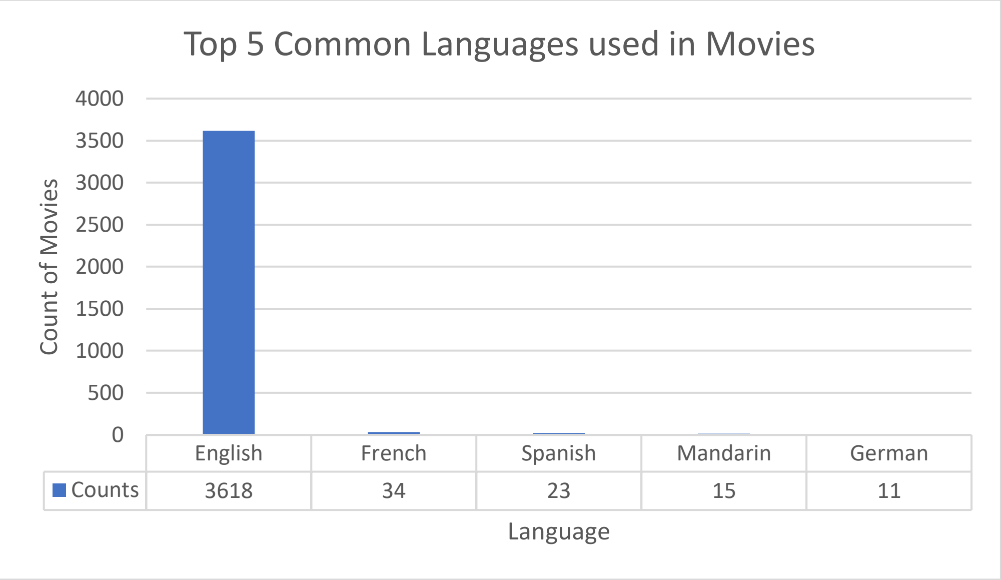


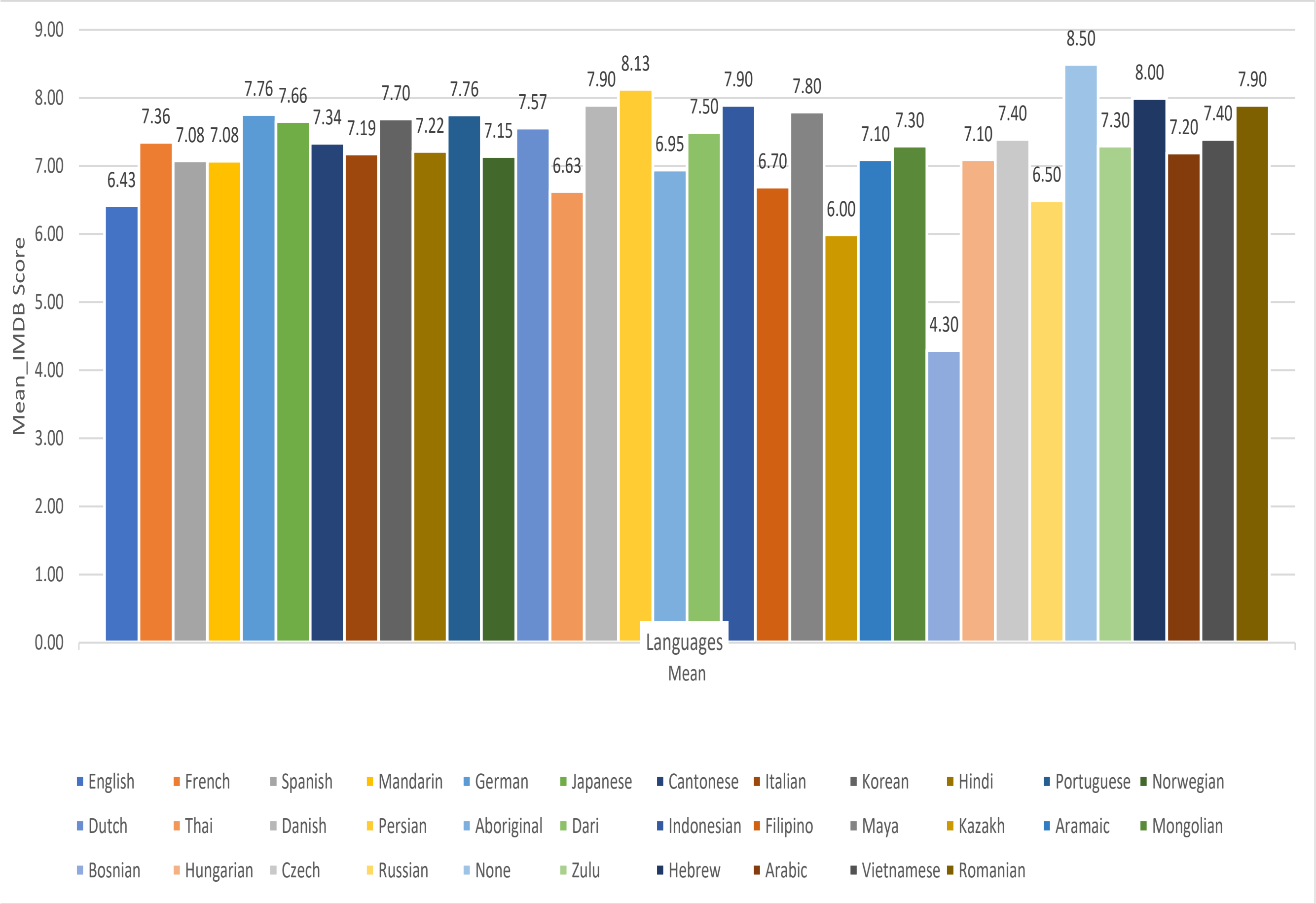


Q No. 3	C. Language Analysis: Situation: Examine the distribution of movies based on their language. Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics. Hint: Use Excel's COUNTIF function to count the number of movies for each language. Calculate the mean, median, and standard deviation of the IMDB scores for each language. Compare the statistics to understand the impact of language on movie ratings.

Language	Counts	Mean	Median	Stdev
English	3618	6.43	6.50	1.05
French	34	7.36	7.30	0.52
Spanish	23	7.08	7.20	0.86
Mandarin	15	7.08	7.40	0.77
German	11	7.76	7.80	0.68
Japanese	10	7.66	8.00	0.99
Cantonese	7	7.34	7.30	0.35
Italian	7	7.19	7.00	1.16
Korean	5	7.70	7.70	0.57
Hindi	5	7.22	7.40	0.80
Portuguese	5	7.76	8.00	0.98
Norwegian	4	7.15	7.30	0.57
Dutch	3	7.57	7.80	0.40
Thai	3	6.63	6.60	0.45
Danish	3	7.90	8.10	0.53
Persian	3	8.13	8.40	0.55
Aboriginal	2	6.95	6.95	0.78
Dari	2	7.50	7.50	0.14
Indonesian	2	7.90	7.90	0.42
Filipino	1	6.70	6.70	#DIV/0!
Maya	1	7.80	7.80	#DIV/0!
Kazakh	1	6.00	6.00	#DIV/0!
Aramaic	1	7.10	7.10	#DIV/0!
Mongolian	1	7.30	7.30	#DIV/0!
Bosnian	1	4.30	4.30	#DIV/0!
Hungarian	1	7.10	7.10	#DIV/0!
Czech	1	7.40	7.40	#DIV/0!
Russian	1	6.50	6.50	#DIV/0!
None	1	8.50	8.50	#DIV/0!
Zulu	1	7.30	7.30	#DIV/0!
Hebrew	1	8.00	8.00	#DIV/0!
Arabic	1	7.20	7.20	#DIV/0!
Vietnamese	1	7.40	7.40	#DIV/0!
Romanian	1	7.90	7.90	#DIV/0!

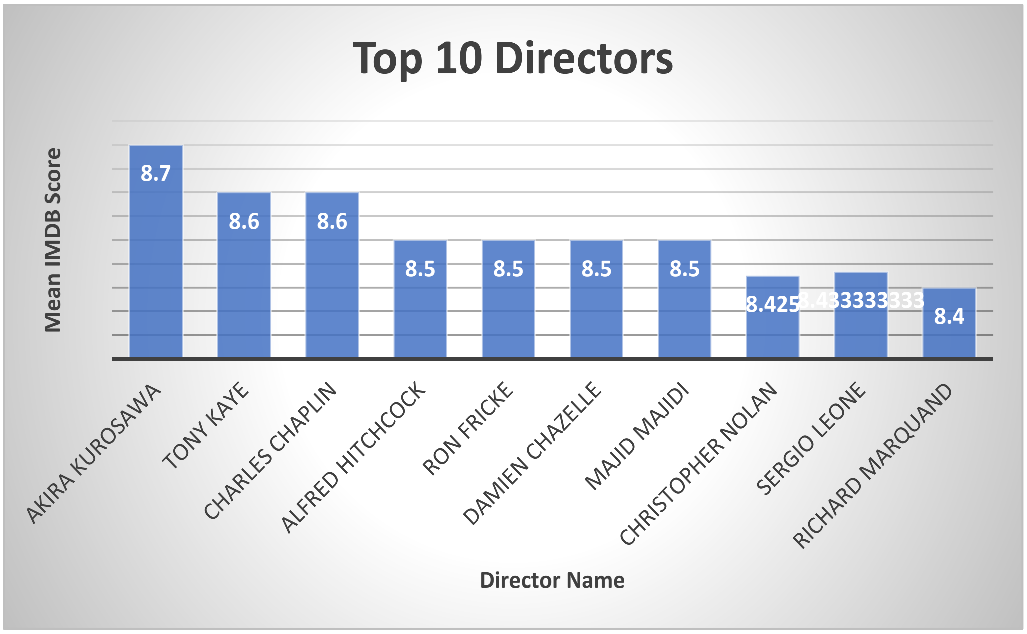
contd.


Top 5 Common Languages used in Movies
Count of Movies
4000
3500
3000
2500
2000
1500
1000
500
0
English
French
Spanish
Mandarin
German
Counts
3618
34
23
15
11
Language


Mean_IMDB Score
9.00
8.00
7.00
6.00
5.00
4.00
3.00
2.00
1.00
0.00
6.43
7.36
7.08
7.08
7.76
7.66
7.34
7.19
7.70
7.22
7.76
7.15
7.57
6.63
7.90
8.13
6.95
7.50
7.90
6.70
7.80
6.00
7.10
7.30
4.30
7.10
7.40
6.50
8.50
7.30
8.00
7.20
7.40
7.90
Languages
Mean
English
French
Spanish
Mandarin
German
Japanese
Cantonese
Italian
Korean
Hindi
Portuguese
Norwegian
Dutch
Thai
Danish
Persian
Aboriginal
Dari
Indonesian
Filipino
Maya
Kazakh
Aramaic
Mongolian
Bosnian
Hungarian
Czech
Russian
None
Zulu
Hebrew
Arabic
Vietnamese
Romanian

Q No. 4	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. Hint: Calculate the average IMDB score for each director. Use Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores.

Unique Director Names	Mean	Percentile	Comparison
Akira Kurosawa	8.7	1	Above Average
Tony Kaye	8.6	0.998	Above Average
Charles Chaplin	8.6	0.998	Above Average
Alfred Hitchcock	8.5	0.996	Above Average
Ron Fricke	8.5	0.996	Above Average
Damien Chazelle	8.5	0.996	Above Average
Majid Majidi	8.5	0.996	Above Average
Christopher Nolan	8.425	0.995	Above Average
Sergio Leone	8.433333333	0.995	Above Average
Richard Marquand	8.4	0.994	Above Average



E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

Q No. 5

Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.
Hint: Calculate the correlation coefficient between movie budgets and gross earnings using Excel's CORREL function. Calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin using Excel's MAX function.

	Correlation Coefficient
	0.1003
Movie Name	**Max profit Margin (G-B)**
Avatar	523505847
Movie Name	**Max profit Margin (%)**
Paranormal Activity	719348.55%

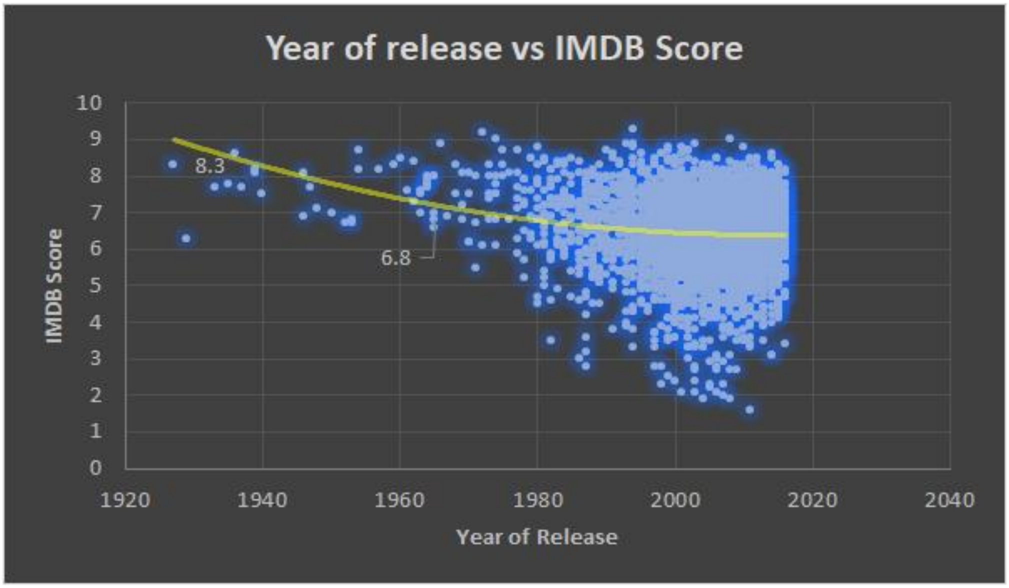
Insights

- **Genre Distribution**: - there are two ways of doing this particular analysis, there are two types of genres, one is individual and the other is hybrid genres, I opted to work on individual genre.
    - It is evident from the observations that Drama is the most common Genre in movies followed by comedy and thriller
    - While statistics suggest that “Film-noir” has the highest average Imdb_score but ther’s only one movie in that category(as we can see the table it has no Mode Variance and Standard deviation), the next highest is History with 7.148 Mean IMDB score and it falls under 150 movies from the dataset, and third is biography with 7.142 Mean IMDB score and it falls under 242 movies from the dataset.
    - Top 5 most common Genre’s are Drama, Comedy, Thriller, Action, Romance in its respective order.
- **Movie Duration Distribution**: - most of the movies falls under 98minutes to 114 minutes in this data set , looking at the statistics of the duration Mean is 110.228 and median is 106 which falls under the above stated range,
    - It is observed from the scatterplot that as the duration of movies increase the volume of movies is decreased which means there are limited movies with higher than 250 minutes duration in comparison to the dataset and IMDB Score seemed to increase with duration and at certain point it started dropping it which is an indication of a range of duration that is good for considering duration of movies.
- **Language Analysis**: - it is very clear and evident from the observation that English language is dominating in most of movies (i.e over 95.79%) , followed by French and Spanish with movie counts of 34 and 23 respectively.
    - As seen in the chart by IMDB Score, if we do not consider the “None” category of language, the highest Average IMDB score of 8.13 is on “Persian” language followed by “Hebrew” language having 8.00 Average IMDB Score with movie counts 3 and 1 respectively.
    - Least average IMDB score of 6.0 is with Kazakh on 1 movie count
- **Director Analysis**: - as observed from the leaderboard like table of Directors, top 10 directors ranked by Mean IMDB score ranges between 8.4 to 8.7 having 10th director on 99.4 percentile
    - Top 10 directors were compared with their respective Mean IMDB Scores by overall mean IMDB Scores and all top 10 are evidently above average.

  - Having a good director indeed leads the entire crew in a systemized manner and operations of that projects would be smooth given high experience of the said director.

- **Budget Analysis**: -
  - The Profit of Movies are calculated with Gross-Budget and the highest profit-making movie in the entire data set is “Avatar” (i.e 523505847)
  - Correlation coefficient is relatively poor here suggesting poor relation between the budget and gross earning which ultimately suggests that having more budget is not sufficient for success of movie
  - The word profit margin basically means finding percentage of the margins achieved, hence the best profit margin in % is “Paranormal Activity” that movie had a budget of $ 15000 only and it ended up making $107917283 giving a robust 719348.55% profit margin an unimaginable return.

Extra #1	**Visualize a relationship between Year of Release and its IMDB Score**



Based on the analysis of the scatter plot depicting the relationship between movie release year and IMDB scores, several noteworthy insights emerge.

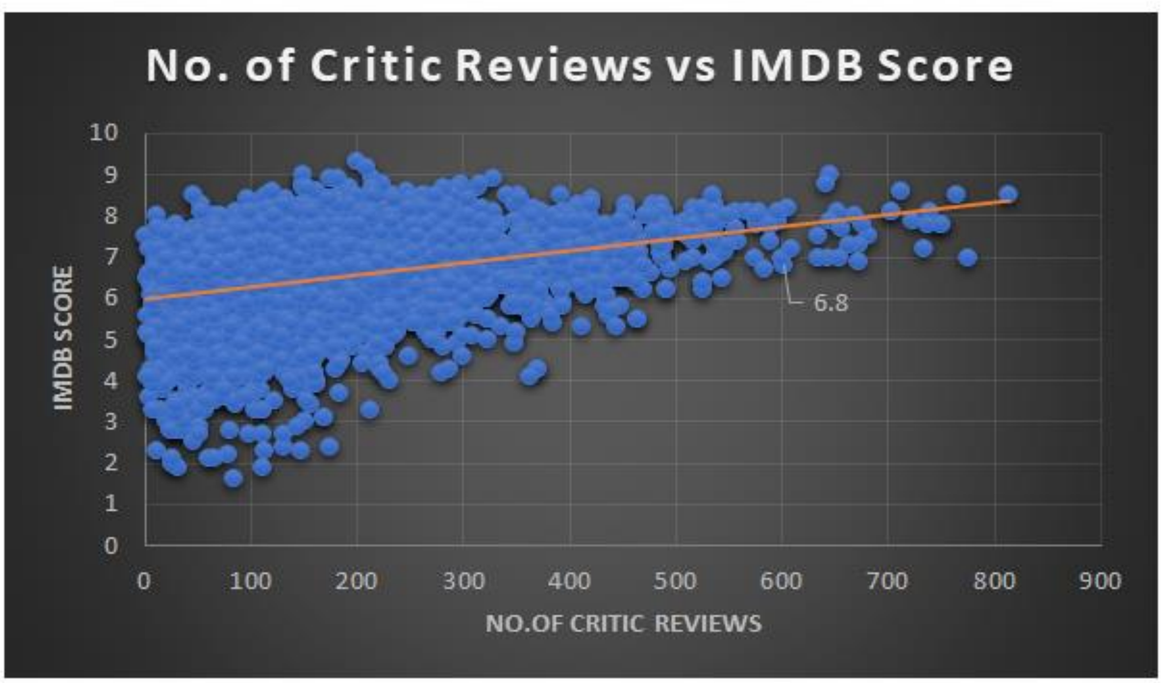
Initially, spanning from 1927 to 1966, a notable consistency is observed in the IMDB scores of movies, which remained relatively stable within a range of 6.8 to 8.9. During this period, the cinematic landscape likely featured a more limited number of releases, potentially allowing for a more discerning audience and stricter quality standards.

However, a discernible shift occurs post-1966, characterized by a substantial increase in the volume of movie releases. This surge in cinematic production coincides with a noticeable broadening of the IMDB score range, expanding from 1.6 to 9.3.

This widening spectrum of IMDB scores suggests a diversification in cinematic offerings, with films spanning a broader range of genres, themes, and artistic merits. Additionally, it may reflect evolving audience tastes, as well as the democratization of filmmaking tools and platforms, enabling a greater diversity of voices and narratives to enter the cinematic landscape.

Overall, these observations underscore the dynamic nature of the film industry over time, marked by shifts in both quantity and quality of movie releases, reflecting the evolving preferences and cultural dynamics of audiences worldwide.

Extra #2 Visualize a relationship betweenNo. Of critic reviews and its IMDB Score



Upon analyzing the scatter plot depicting the relationship between the number of critic reviews and IMDB scores, several insightful observations come to light.

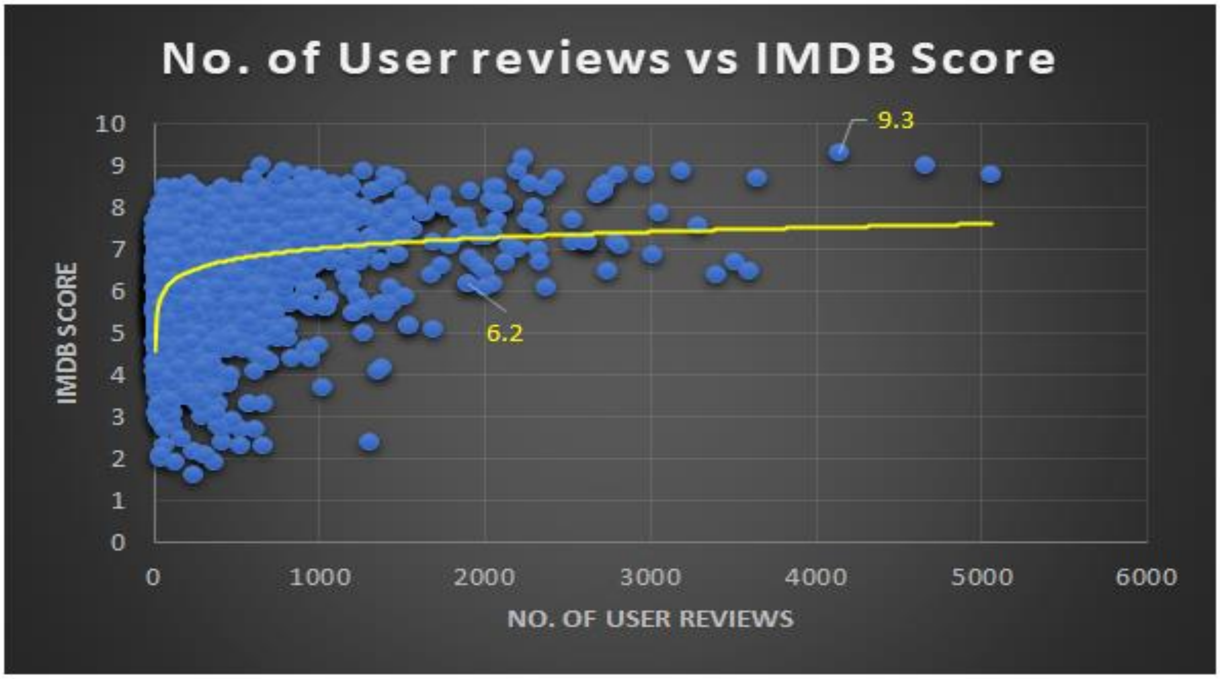
Initially, it becomes apparent that movies with a lower number of critic reviews exhibit a wide range of IMDB scores. This variance suggests that the quantity of critic reviews does not exert a significant influence on the IMDB score within this range.

However, as the number of critic reviews increases, a discernible trend emerges. Specifically, there is a noticeable uptick in IMDB scores corresponding to higher numbers of critic reviews. Notably, movies with over 600 critic reviews consistently attain IMDB scores above 6.8. This pattern suggests a positive correlation between the volume of critic reviews and IMDB scores, implying that a greater number of critic reviews tends to elevate the perceived quality of a movie.

This finding underscores the importance of critical reception in shaping the perceived quality and success of a movie. As the number of critic reviews serves as a proxy for the level of attention and scrutiny a movie receives, a higher volume of reviews can enhance audience trust and confidence in a film's merits, thereby positively impacting its IMDB score.

In conclusion, the observed correlation between the number of critic reviews and IMDB scores highlights the significance of critical acclaim in influencing audience perceptions and contributing to the overall reception of a movie within the cinematic landscape.

Extra #3	Visualize a relationship betweenNo. Of user reviews and its IMDB Score



In examining the scatter plot depicting the correlation between the number of user reviews and IMDB scores, a series of insightful patterns emerges.

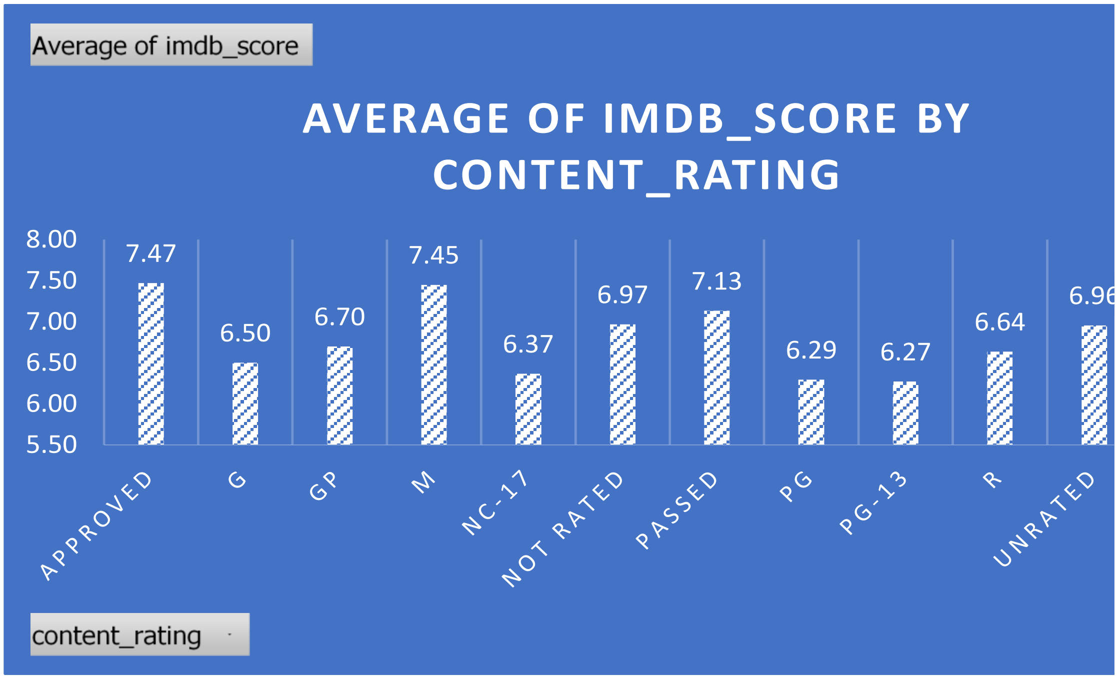
Initially, the plot reveals a broad spectrum of IMDB scores associated with movies receiving a lower number of user reviews. Within this range, the impact of user feedback on IMDB scores appears to be relatively minimal.

However, as the quantity of user reviews increases, a distinct trend begins to manifest. Notably, there is a noticeable uptick in IMDB scores as the number of user reviews rises. Particularly noteworthy is the consistent performance of movies with over 1900 user reviews, which consistently achieve IMDB scores surpassing 6.2.

This consistent pattern underscores a positive correlation between the volume of user reviews and the perceived quality of movies. Consequently, it suggests that a higher volume of user feedback tends to elevate a movie's IMDB score, reflecting positively on its reception among audiences.

These observations underscore the significance of user engagement and feedback in shaping audience perceptions and influencing the overall reception of a movie. As such, the findings emphasize the pivotal role of user reviews in shaping the success and reputation of movies within the cinematic landscape.

Extra #4	**Write insight on efeects of content rating on Avg IMDB Score**



After closely analyzing the bar chart depicting content ratings alongside their respective mean IMDB scores, significant patterns emerge, shedding light on how content classification impacts audience reception.Remarkably, movies classified under the "Approved" and "M" (Mature) ratings exhibit a notably higher average IMDB score, standing at an impressive 7.45 and above. This trend suggests a consistent preference among audiences for content within these categories, characterized by mature themes and content.

Conversely, movies categorized as "PG-13" and "PG" (Parental Guidance Suggested) demonstrate a comparatively lower average IMDB score, hovering around 6.29 and below. These ratings typically indicate content deemed suitable for older adolescents or those requiring parental guidance due to potentially inappropriate material for younger audiences.

The disparity in average IMDB scores between these rating categories underscores a distinct preference among audiences for content aligned with the "Approved" and "M" ratings, reflecting a higher likelihood of achieving a commendable IMDB score.

This observation underscores the critical role of content rating in shaping audience perceptions and preferences, highlighting the significance of aligning content with audience expectations to optimize viewer satisfaction and reception within the cinematic landscape.

Results

- A. Drama is the most common genre followed by comedy, thriller, action, romance
- B. 110.228 and 106 are mean and median of duration column in minutes and highest number of movies falls under 98 to 114 minutes bracket
- C. English dominates all movies in this dataset over 95% of movies are in English and movies with higher mean IMDB scores are Persian and Hebrew
- D. Director with highest mean IMDB score of 8.7 is Akira Kurosawa.
- E. Avatar is the highest Profit-making movie and Paranormal activity has the highest Profit margin (%).

Click here for Excel sheet

Click here for Video Presentation

www.linkedin.com/in/kartikey-vyas