IMDB Movie Analysis

- Report by – Siva Sankari H

1.Introduction

IMDB analysis refers to the process of examining and extracting from data on IMDb (*Internet Movie Database)*, which is a popular online platform providing information about movies, TV shows, actors, directors, and other related content. The analysis typically focuses on various aspects of the data to identify patterns, trends, and make informed decisions based on that data.

In our project we mainly focus on the factors that affects the movie success. Here success is defined by high ratings. IMDb provides user-generated ratings for movies, TV shows, and other media. Analysing ratings can reveal the popularity, critical reception, and audience preferences over time. Researchers might study how ratings change over time, or how different genres perform relative to others.

2.Project Description

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

2.2 Dataset –

We are going to use the dataset provided along with the question. We need to perform data cleaning and data analysis. Cleaning of data involves – preprocessing, handling missing values, removing duplicates, converting data types if necessary and possibly feature engineering. In analysis part we will explore the data to understand the relationships between different variables. We will look at the correlation between movie ratings and other factors like genre, director, budget, etc.

Data Cleaning –

1. There are 45 duplicates. We have removed it. The dataset now has 4998 unique values. We can find blanks in most of the columns. We are going to remove it according to the analysis performed.
2. We are removing unwanted columns from our dataset in order to facilitate easy analysis. Out of the available fields we will only use columns having information about director, movie title, duration, gross, genre, budget, imdb score, language, and country.

Dataset Analysis –

- Genre category contains 26 unique values. All the movies in our dataset falls under at least one of these genres.

- There are no blanks in director column. The duration, gross, budget column has some blanks which will be removed.

Analysis Performed in this project – Movie Genre Analysis, Movie Duration Analysis, Language Analysis, Movie Budget analysis and Director Analysis.

3.Apporach, Analysis and Insights Achieved

A. Movie Genre Analysis – Analyse 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.

Approach –

Excel Function Used – text-to-column conversion for splitting the genre column. Excel's COUNTIF function to count the number of movies for each genre. Functions like AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV are used to calculate descriptive statistics.

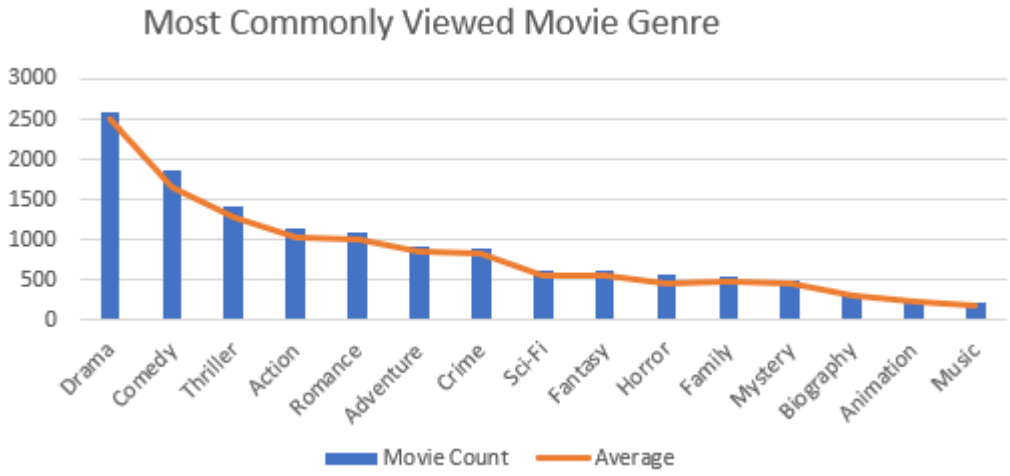
Analysis –

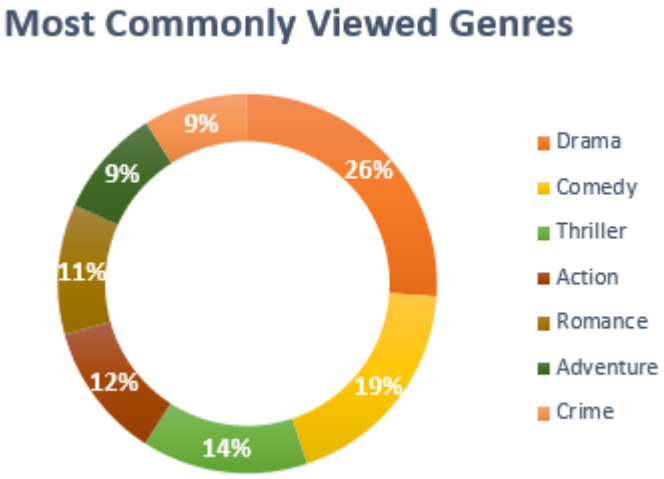
We have a total of 26 genres. All the categories are combined in one column. We need to split it and perform analysis. There is a maximum of 7 genres tagged for movie with least being 1. We need to identify the most common movie genre. We have a total of 26 genres. All the movies fall under any one of these genres. The plot below shows the distribution highlighting top 15 categories. Observations –

- Main stream genres are Drama, Comedy, Action, Romance, and thriller. The audiences love for movies decreases as we move down towards more distinct genres.
- Movies outside the mainstream genres often struggle to find a large audience, leading to lower box office success.
- The average viewers rate also decrease as we move towards non main streams genres.

Genre	Movie Count	Average
Drama	2594	2506.457
Comedy	1872	1656.786
Thriller	1408	1270.371
Action	1153	1027.800
Romance	1106	1019.100
Adventure	923	849.314
Crime	889	833.729
Sci-Fi	616	552.800
Fantasy	610	549.614
Horror	565	471.657
Family	546	487.114
Mystery	500	463.314
Biography	293	299.286

Animation	242	227.343
Music	214	195.971





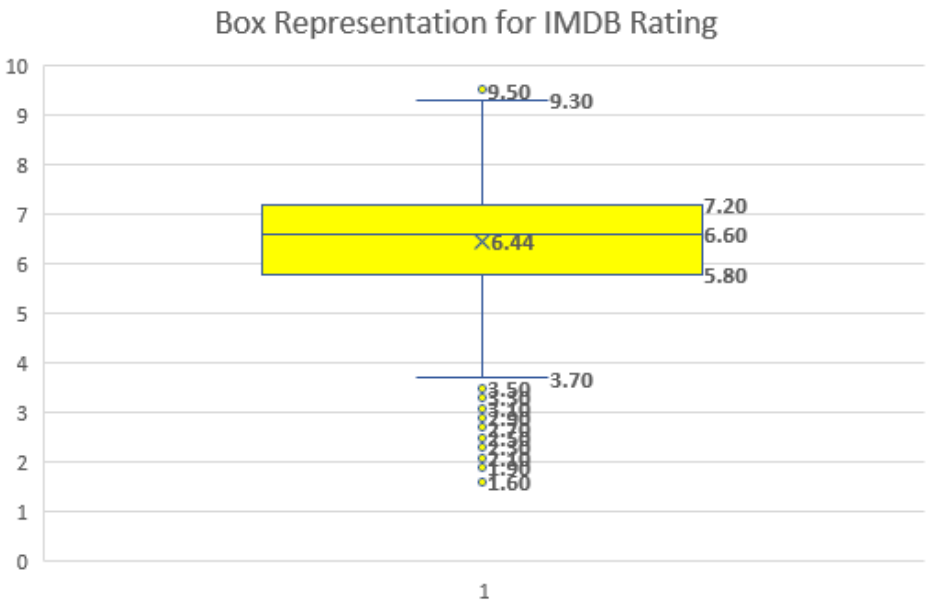
B. Movie Duration Analysis: Analyse the distribution of movie durations and its impact on the IMDB score.

Task: Analyse the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Approach and Analysis –

Excel Function Used – Excel's functions like AVERAGE, MEDIAN, and STDEV to calculate statistical measures. Scatter plot to visualize the relationship between movie duration and IMDB score and trendline to assess the direction and strength of the relationship.

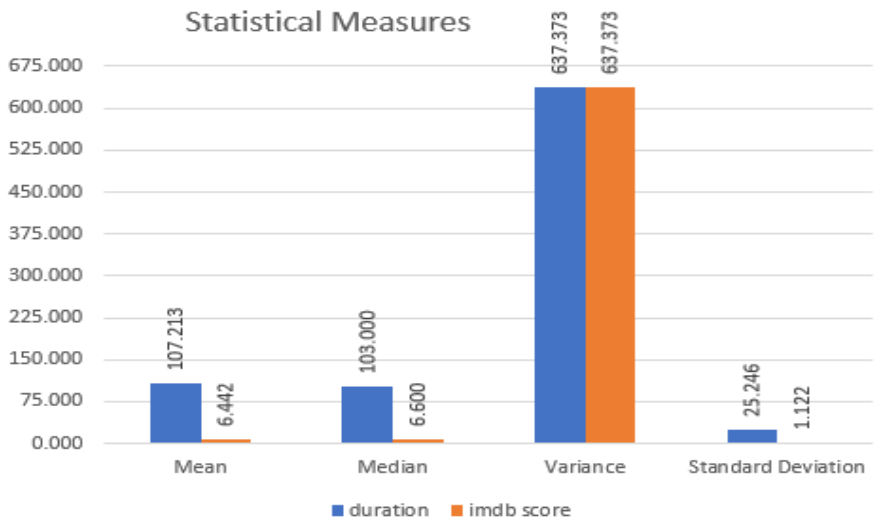
Since duration plays a major role in this analysis we are eliminating rows with no duration. Let's analyse how the imdb score is spread out.

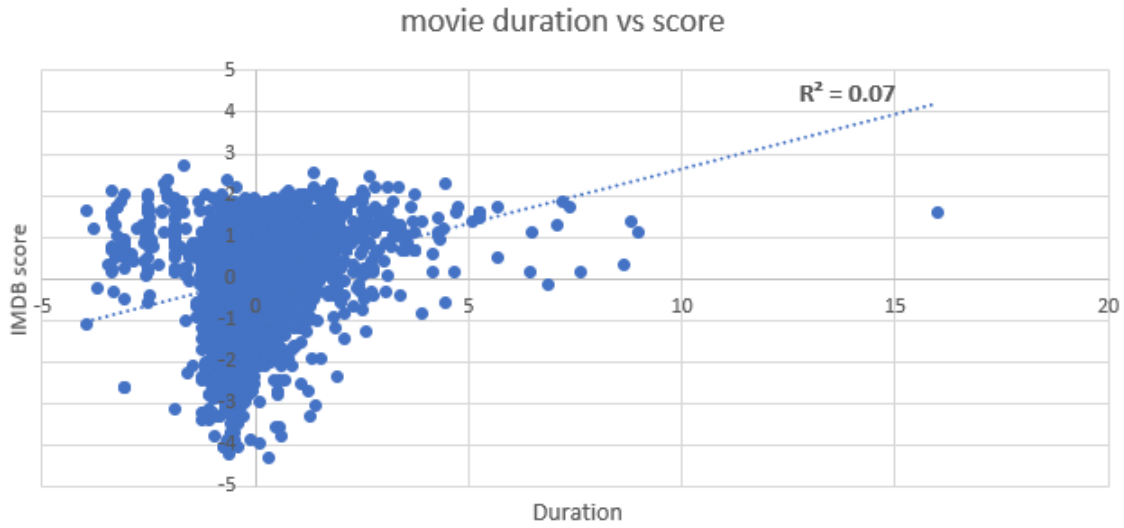


Box plot Chart analysis for imdb score for all the movies is as follows –

- Mean = 6.44. most of the data is centred around this point. The distribution is slightly skewed to the left. This is suggested by the longer whisker extending downward compared to the upper whisker.
- The median = 6.6 falls slightly above the mean
- Lower quartile = 5.8 and upper quartile = 7.2.
- First 25% of the score lies below 5.8 rating with rating 3.7 being the minimum value. The points below the minimum line represents the outliers.
- 75% of the score lies below 7.2 rating.

Statistical Analysis –





Observation of the scatter plot –

- Weak correlation – we have a weak correlation between duration and imdb score. R^2 = 0.07 indicates the same. This means only 7% of the scores can be explained by the duration. The low R^2 value suggests that movie duration is not a good predictor of score. Other factors, such as genre, plot, acting, and directing, likely play a much bigger role in determining a movie's score than its duration.
- Majority of the points are scattered around the trend line and not on it – indicating weak relationship.
- The slope of the trend line is itself not very steep showing poor relationship.
- R^2 value can be improved if we move towards polynomial trendline. A maximum R^2 value of 0.25 can be obtained with a polynomial of order 5. Only disadvantage is as the order of polynomial increase trendline tends to overfit.

C. Language Analysis: 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.

Approach and Analysis –

Excel Function Used – Excel's COUNTIF function to count the number of movies for each language. Mean, Median, and Standard Deviation to calculate statistical measures for each language. Compare the statistics to understand the impact of language on movie ratings.

Observations -

- From the Excel output it can be seen that “English” is the most common language. Around 4585 movies have been made in English.
- The average rating for the same is 6.39. it can be seen that languages English, Hindi, Russian and Italian have very high variance and considerably high mean.
- English movies often receive high ratings on IMDb due to several factors. The large global audience for English-language films leads to more ratings, potentially boosting scores.

- Additionally, the influence of English-speaking countries like the US and UK in the film industry increases exposure and recognition for these movies.
- Finally, Hollywood's high production values and advanced filmmaking techniques contribute to a perception of higher quality, further influencing ratings.

Top 10 Used Languages –

Language	Count	Sum	Mean	Variance	SD	Maximum
English	4585	29315.30	6.39	1.27	1.13	9.50
French	73	513.80	7.04	0.52	0.72	8.40
Spanish	40	277.50	6.94	0.71	0.84	8.20
Hindi	28	185.70	6.63	1.89	1.37	8.50
Mandarin	24	162.90	6.79	1.03	1.02	7.90
German	19	139.50	7.34	0.86	0.93	8.50
Japanese	17	124.90	7.35	0.94	0.97	8.70
Italian	11	79.50	7.23	1.41	1.19	8.90
Russian	11	70.00	6.36	1.74	1.32	8.10
Cantonese	11	76.50	6.95	0.45	0.67	7.80

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.

Approach and analysis –

Excel Function Used – Excel's PERCENTILE function to identify the directors with the highest scores. Compare the scores of these directors to the overall distribution of scores. We will use director, movie name, imdb score information from the dataset to measure the percentile values.

Average IMDB values for each director can be calculated by using excel average function.

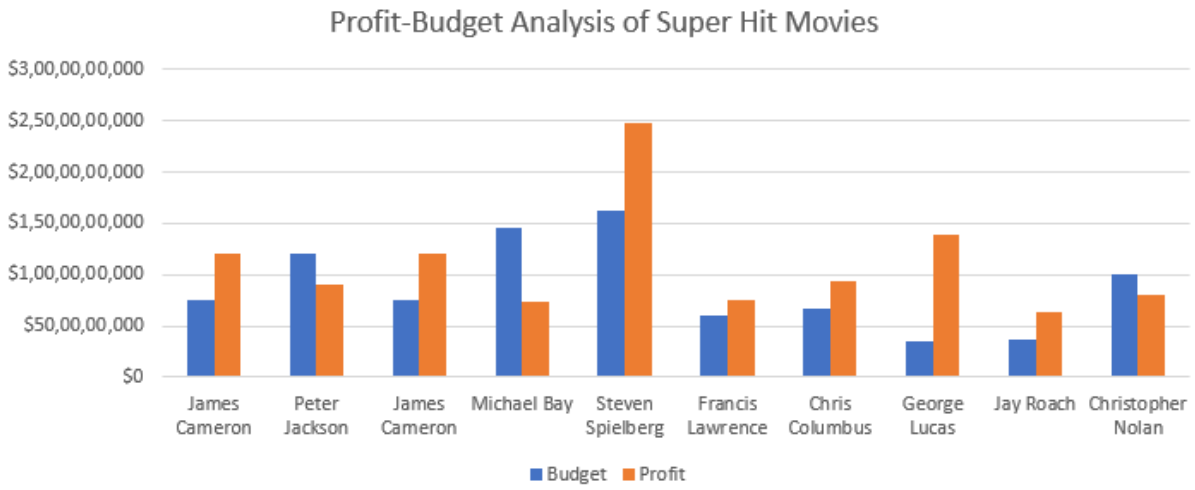
$$avg\ score = AVERAGEIF(Table6[director\_name], V2:V2400, Table6[imdb\_score])$$

We have information of 1754 directors having full information on their imdb scores, movies list and success rate. All these directors have rating ranging from 8.6 maximum to minimum 2.1.

Top 10 scores fall under the 8.6 to 8.4 range. The list of directors falling under this category are as follows –

Director name	Movies Made	Average imdb score	percentile score
Tony Kaye	1	8.600	0.988
Charles Chaplin	1	8.600	0.988
Alfred Hitchcock	1	8.500	0.984

Ron Fricke	1	8.500	0.984
Damien Chazelle	1	8.500	0.984
Majid Majidi	1	8.500	0.984
Sergio Leone	3	8.433	0.981
Christopher Nolan	8	8.425	0.981
S.S. Rajamouli	1	8.400	0.980
Richard Marquand	1	8.400	0.980



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

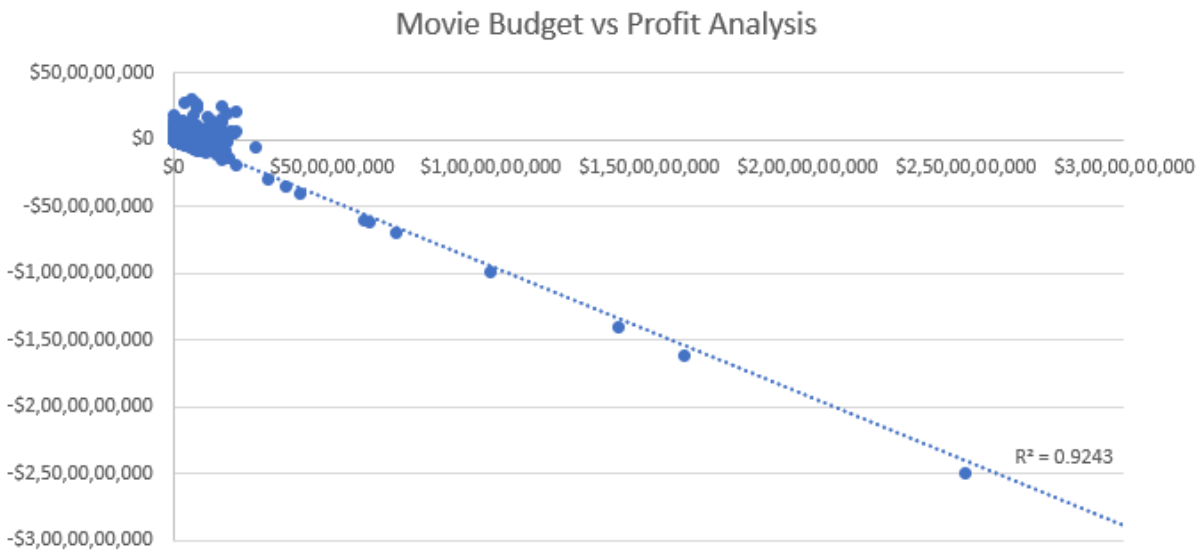
Task: Analyse the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Excel Function Used – Scatter plot used to analyse the spread of budget and profit details.

$$Profit = Gross\ Amount - Movie\ Budget$$

A film is said to be successful if its gross value is greater than the budget of the movie. Several factors affect the profit of the movie. The plot below shows one such factor. We are analysing the relation between budget and profit. Observations are as follows –

- Negative correlation – as movie budget increase it may not necessarily lead to profit. Profit generally decreases.
- R^2 = 0.9243 indicates a strong correlation between the variables. This means 92.4% of the profit can be explained by the budget.
- There are few visible outliers in the plot which is deviating the relation. Even if we remove that outlier we will get a R^2 = 0.85 which further indicates negative correlation.



4.Tech Stack Used

I have used Microsoft Excel to perform this analysis. Excel's function like mean, median, variance standard deviation all are used for descriptive analysis – it has helped in gaining meaningful insights. We have used different excel charts for visualizing the data.

5. Drive Link for Excel

Modified_Excelsheet