

# IMDB Movie Analysis

## Project Description:

The dataset offered comprises information on IMDB Movies, allowing you to explore the elements that influence a movie's performance on the site. In this context, success is characterized by high IMDB ratings, which are a reliable indicator of a film's popularity and quality. We want to identify the primary elements influencing a movie's rating by examining numerous parameters such as genre, budget, cast, director, and release date. This study is extremely useful for producers, directors, and investors who want data-driven insights to help them make decisions about future projects. Understanding these elements can assist stakeholders in optimizing their strategy for producing or investing in films with a higher chance of success with audiences.

## Project Approach:

### **Define the Objective:**

The main goal is to identify key factors that influence a movie's success on IMDB, where success is defined by high IMDB ratings. Understanding these factors can help producers, directors, and investors make informed decisions.

### **Data Cleaning and Preparation:**

Start by thoroughly cleaning the dataset. This includes:

1. Handling missing values (e.g., imputing or removing them).
2. Removing duplicates to ensure accurate results.
3. Converting columns to appropriate data types (e.g., numerical or categorical).
4. Feature engineering (e.g., creating new variables like "budget range" or "release decade" if needed).

### **Exploratory Data Analysis (EDA):**

Perform an initial exploration of the data to understand its structure and key trends:

1. Analyze the distribution of variables like genre, duration, and language.
2. Use visualizations (histograms, box plots, etc.) to understand relationships between factors such as budget, genre, and IMDB scores.
3. Identify outliers and trends that might influence the analysis.

### **Genre Analysis:**

Investigate how different genres impact IMDB ratings:

1. Calculate descriptive statistics (mean, median, variance) for each genre.
2. Compare how certain genres tend to perform in terms of ratings.

### **Duration Analysis:**

Analyze how the length of a movie affects its rating:

1. Calculate the average, median, and standard deviation of movie durations.
2. Create scatter plots to visualize the relationship between movie duration and IMDB ratings.

### **Language Analysis:**

Examine whether certain languages have a significant impact on movie ratings:

1. Identify the most common languages in the dataset.
2. Use descriptive statistics to compare IMDB scores for each language.

### **Director Analysis:**

Analyze the influence of directors on movie ratings:

1. Calculate the average IMDB score for each director.
2. Use percentile calculations to highlight top-performing directors and compare them with overall rating distributions.

### **Budget and Profit Margin Analysis:**

Explore the relationship between a movie's budget and its financial success:

1. Calculate the correlation between budget and gross earnings.

2. Identify the movies with the highest profit margins and assess whether higher-budget movies consistently have better ratings.

### Five Whys Approach:

Apply the "Five Whys" technique to dive deeper into insights. For example, if higher-budget movies have higher ratings, ask why this occurs to uncover underlying reasons like production quality or marketing impact.

The tech stack I used is **MS EXCEL 2016**

## Insights:

### Task1:


**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.

- ❖ \_To do this work, I separated the genre column using Google Sheets' text\_to\_columns option. Then, using Excel tools, I deleted duplicate Genres and calculated the number of movies for each genre, as well as the average, median, mode, max, min, variance, and standard deviation.

### Result:

Average	219
Median	37
MAX	984
MIN	1
Mode	1
Variance	110065.9
Stand Dev	331.7618

Genres	 Count of genres
Action	951
Adventure	366
Animation	45
Biography	204
Comedy	984
Crime	253
Documentary	26
Drama	659
Family	3
Fantasy	37
Horror	159
Musical   Romance	4
Mystery   Thriller	21
Romance   Sci-Fi   Thriller	1
Sci-Fi   Thriller	7
Thriller	1
Western	2
<b>Grand Total</b>	<b>3723</b>

## Task2:

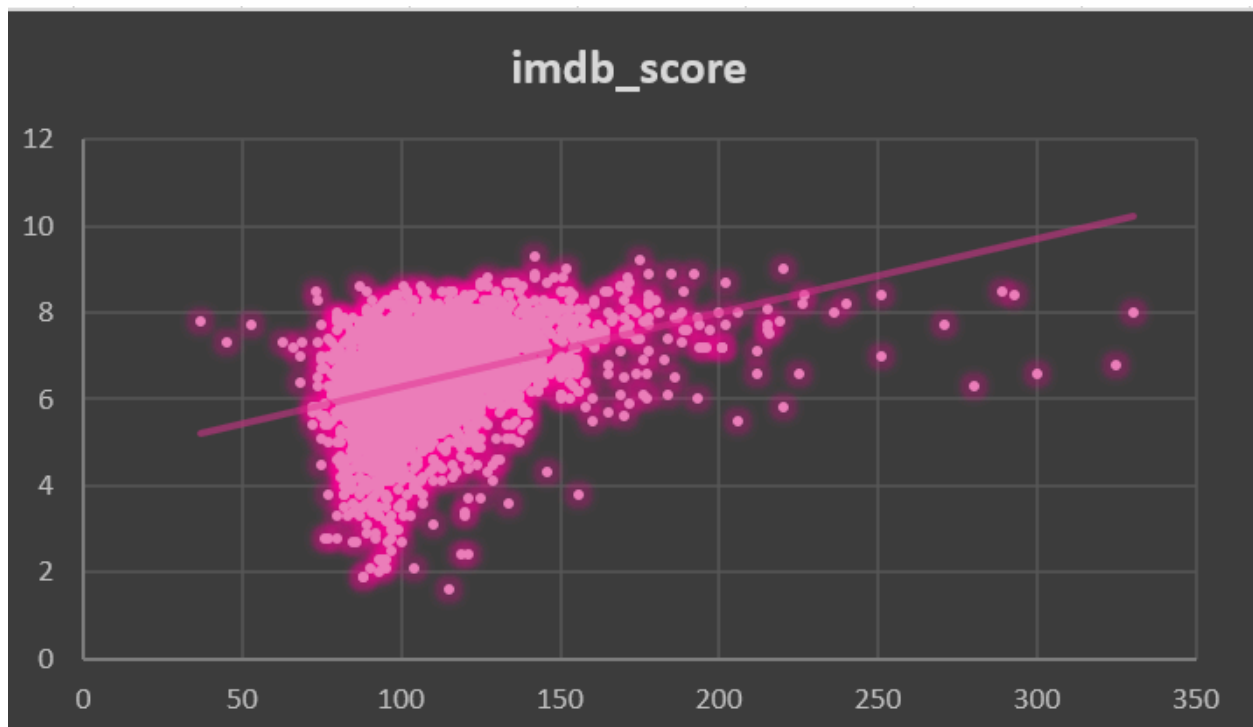
**Movie Duration Analysis:** Analyze the distribution of movie durations and its impact on the IMDB score.

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

- ❖ For this work, I simply used the needed columns, Duration and IMDB\_score. Thus making it easier to perform the work. We may determine the connection among the two columns using the mean, median, and standard deviation.

## Result:

Average	110.2635
Median	106
Stand Dev	22.678325



### Task3:

**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.

### Result:

Row Labels	Count of language	MEAN	MEDIAN	STDEV
Aboriginal	2			
Arabic	1	6.95	6.6	1.05
Aramaic	1	7.20	6.6	1.05
Bosnian	1	7.10	6.6	1.05
Cantonese	7	4.30	6.6	1.05
Czech	1	7.34	6.6	1.05
Danish	3	7.40	6.6	1.05
Dari	2	7.90	6.6	1.05
Dutch	3	7.50	6.6	1.05
English	3566	7.57	6.6	1.05
Filipino	1	6.43	6.6	1.05
French	34	6.70	6.6	1.05
German	10	7.36	6.6	1.05
Hebrew	1	7.77	6.6	1.05
Hindi	5	8.00	6.6	1.05
Hungarian	1	7.22	6.6	1.05
Indonesian	2	7.10	6.6	1.05
Italian	7	7.90	6.6	1.05
Japanese	10	7.19	6.6	1.05
Kazakh	1	7.66	6.6	1.05
Korean	5	6.00	6.6	1.05
Mandarin	14	7.70	6.6	1.05
Maya	1	7.02	6.6	1.05
Mongolian	1	7.80	6.6	1.05
None	1	7.30	6.6	1.05
Norwegian	4	8.50	6.6	1.05
Persian	3	7.15	6.6	1.05
Portuguese	5	8.13	6.6	1.05
Romanian	1	7.76	6.6	1.05
Russian	1	7.90	6.6	1.05
Spanish	23	6.50	6.6	1.05
Thai	3	7.08	6.6	1.05
Vietnamese	1	6.63	6.6	1.05
Zulu	1	7.40	6.6	1.05
Grand Total	3723	7.30	6.6	1.05

By Statistical Analysis, The most common language used in movies is English with 3566 movies. As descriptive statistics we calculated mean, median and standard deviation.

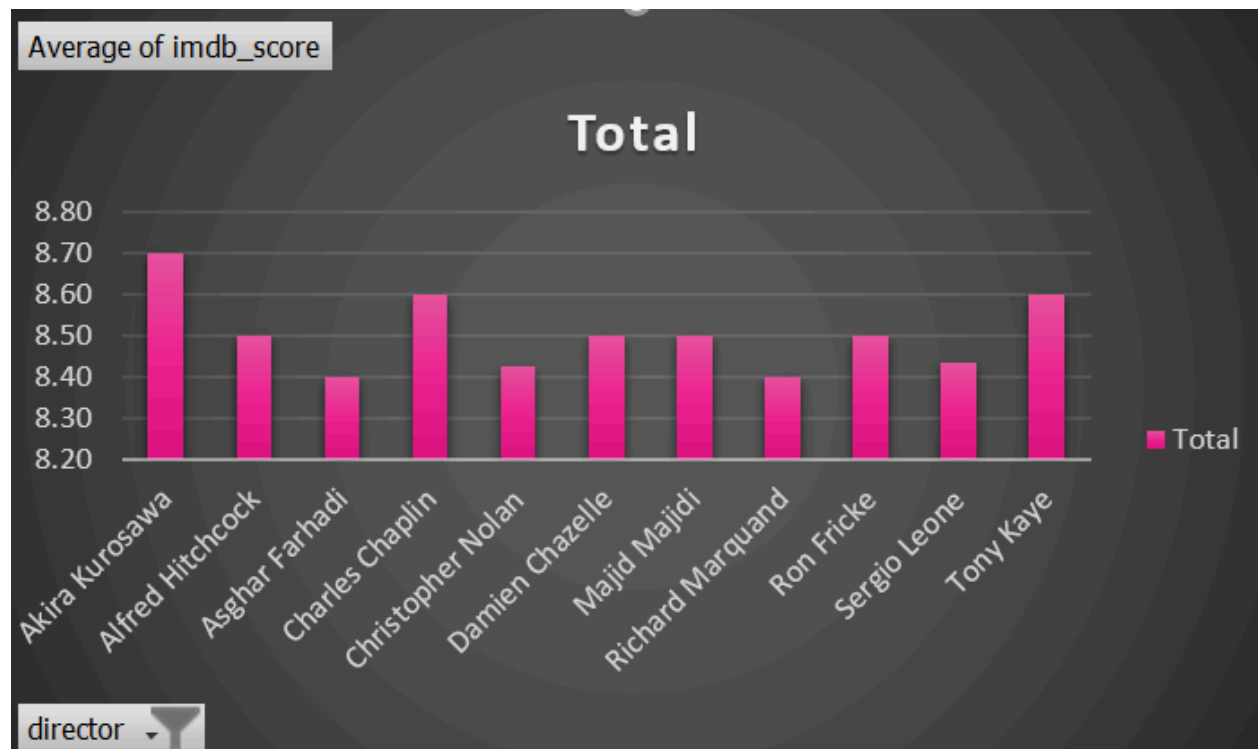
#### Task4:

**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:

Director Name	Average of imdb_score
Akira Kurosawa	8.70
Alfred Hitchcock	8.50
Asghar Farhadi	8.40
Charles Chaplin	8.60
Christopher Nolan	8.43
Damien Chazelle	8.50
Majid Majidi	8.50
Richard Marquand	8.40
Ron Fricke	8.50
Sergio Leone	8.43
Tony Kaye	8.60
Grand Total	8.47



### Task 5:

**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.

- ❖ For this work, I applied specific columns such as movie title, gross, and budget. I calculated the profit margin using the formula  $\text{Profit Margin} = \text{Gross} - \text{Budget}$ . To clean the data with blank cells, I deleted the rows that contained both gross and budget blank cells.

### Result:

CORRELATION	0.1016197
MAX PROFIT MARGIN	523505847

Google Sheets link:

 IMDB Movie Analysis.xlsx