IMDB Movie Analysis Project – Sarthak Rasal

Project Description

Problem Statement:

The dataset provided is related to IMDB Movies. A potential problem to investigate is:
"What factors influence the success of a movie on IMDB?"
Success is defined by high IMDB ratings. This analysis is important for movie producers, directors, and investors to make informed decisions based on the attributes that contribute to a movie's success.

Approach

Data Cleaning:

- Handle missing values.
- Remove duplicates.
- Convert data types where necessary.
- Perform feature engineering if needed (e.g., splitting multi-genre fields).

Data Analysis:

- Explore relationships between IMDB ratings and other features such as genre, duration, director, budget, language, etc.
- Use Excel for descriptive statistics, visualizations, and correlation analyses.
- Apply the Five Whys technique to explore deeper insights.

Five 'Whys' Approach

Example:

- **Q**: Why do movies with higher budgets tend to have higher ratings?
  A: They can afford better production quality.
- **Q**: Why does better production quality lead to higher ratings?
  A: It enhances the viewer's experience.

- **Q**: Why does an enhanced viewer experience lead to higher ratings?
  A: Viewers are more likely to rate a movie highly if they enjoyed watching it.
- **Q**: Why are viewers more likely to rate a movie highly if they enjoyed watching it?
  A: Positive experiences lead to positive reviews.
- **Q**: Why do positive reviews matter?
  A: They influence other viewers' decisions to watch the movie, increasing popularity and success.

I began by thoroughly examining the dataset to understand the structure and purpose of each column. The dataset initially contained 28 columns and 5043 rows. However, I observed that several columns were irrelevant, contained null values, or were entirely blank. Therefore, I proceeded with a comprehensive data cleaning process.

1. **Column Reduction:**
   I removed all columns that were not relevant to the project or failed to offer meaningful insights. After this step, I retained only 9 essential columns: **Director Name**, **Duration**, **Movie Title**, **Genre**, **Budget**, **Gross**, **IMDB Rating**, **Language**, and **Country**.
2. **Removing Blank Rows:**
   I identified blank rows using Excel's "Find & Select" feature by choosing "Go To Special" and then selecting "Blanks." This highlighted all blank entries. Using the shortcut **Ctrl + –**, I selected "Entire Row" to delete all blank rows from the dataset.
3. **Removing Duplicate Rows:**
   Finally, I eliminated any duplicate records to ensure data integrity.

After completing these steps, the dataset was refined to contain **9 columns and 3786 rows**. The cleaned version of the dataset is now ready for analysis.

https://docs.google.com/spreadsheets/d/1DkFyyMnUidH0SQbbJE2f8iFGBe0G68zb/edit?usp=sharing&ouid=117617232609274675030&rtpof=true&sd=true

---

Tech-Stack Used

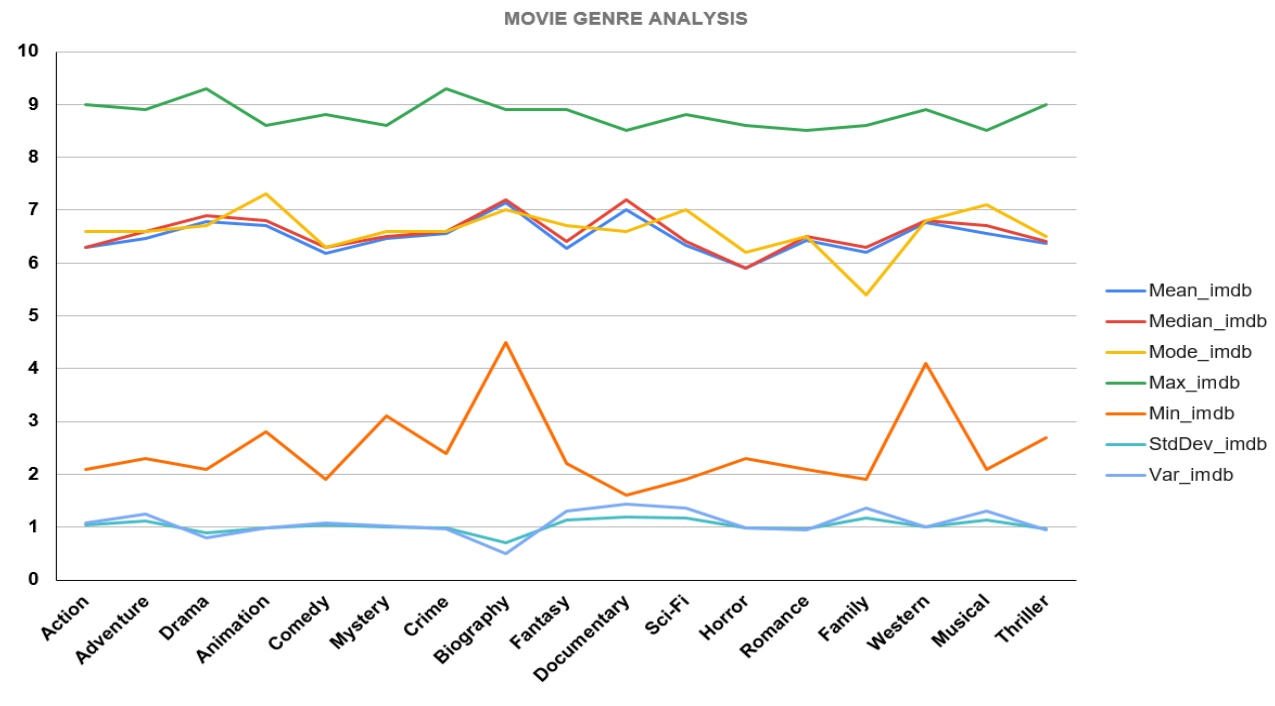
- **Microsoft Excel 2022:**
  Used for data cleaning, performing descriptive statistics, visualizations, correlation analysis, and generating final insights.

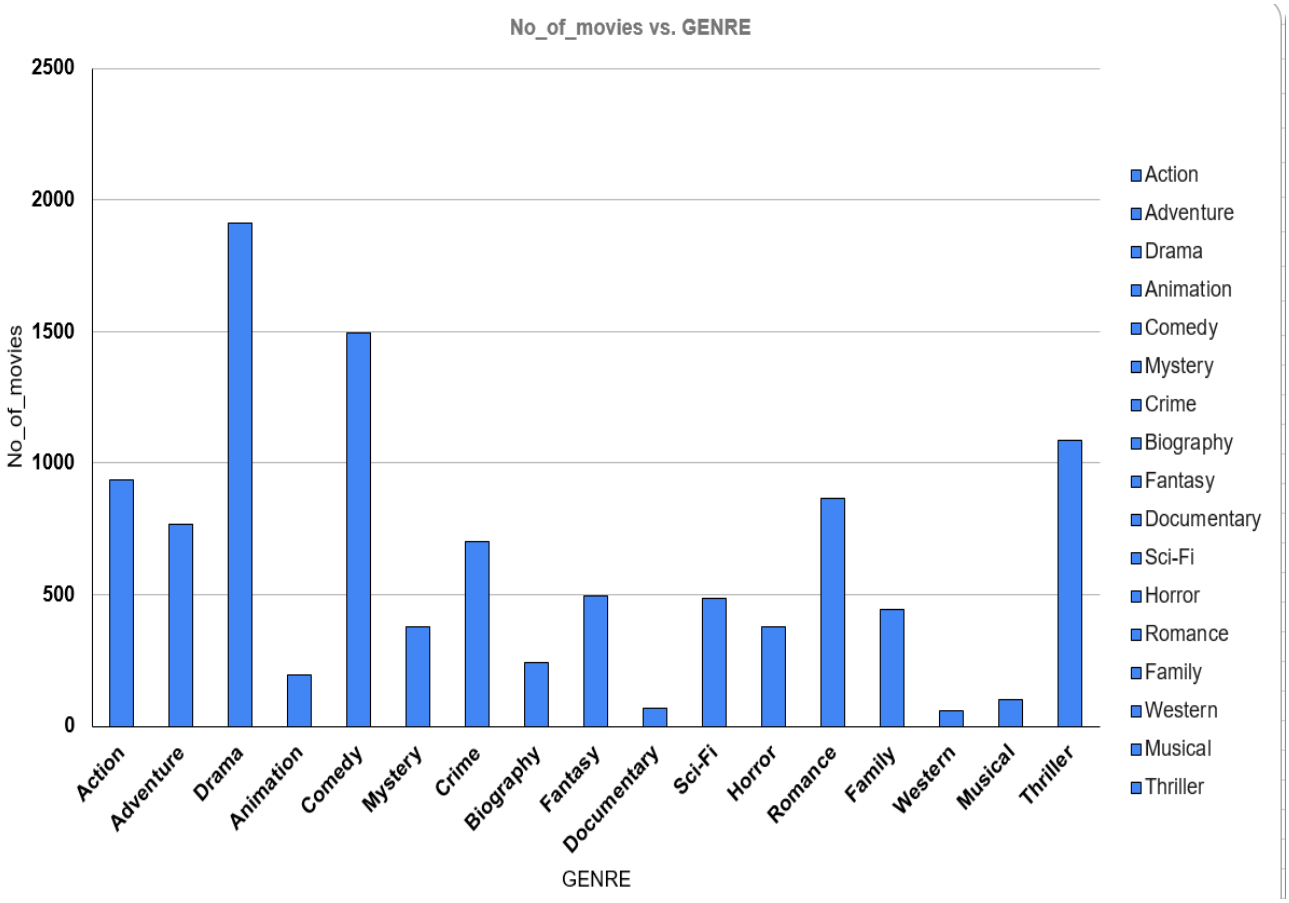
Tasks

A. Movie Genre Analysis

- **Task:** Determine the most common genres and calculate IMDB score statistics per genre (mean, median, mode, range, variance, standard deviation).

GENRE	No_of_movies	Mean_imdb	Median_imdb	Mode_imdb	Max_imdb	Min_imdb	StdDev_imdb	Var_imdb
Action	935	6.285989305	6.3	6.6	9	2.1	1.038357736	1.078186788
Adventure	766	6.454960836	6.6	6.6	8.9	2.3	1.116926308	1.247524378
Drama	1911	6.789115646	6.9	6.7	9.3	2.1	0.891064898	0.793996652
Animation	197	6.700507614	6.8	7.3	8.6	2.8	0.993627526	0.987295659
Comedy	1492	6.183310992	6.3	6.3	8.8	1.9	1.039919012	1.081431552
Mystery	377	6.469496021	6.5	6.6	8.6	3.1	1.007391835	1.014838309
Crime	702	6.548148148	6.6	6.6	9.3	2.4	0.984105199	0.968463042
Biography	242	7.140082645	7.2	7	8.9	4.5	0.71009671	0.504237338
Fantasy	496	6.285080645	6.4	6.7	8.9	2.2	1.140414241	1.30054464
Documentary	67	7.011940299	7.2	6.6	8.5	1.6	1.199939694	1.439855269
Sci-Fi	484	6.327272727	6.4	7	8.8	1.9	1.16718415	1.362318841
Horror	379	5.903957784	5.9	6.2	8.6	2.3	0.991023285	0.982127152
Romance	866	6.426212471	6.5	6.5	8.5	2.1	0.968996249	0.938953731
Family	441	6.2	6.3	5.4	8.6	1.9	1.169576458	1.367909091
Western	58	6.765517241	6.8	6.8	8.9	4.1	0.998516746	0.997035693
Musical	102	6.550980392	6.7	7.1	8.5	2.1	1.143535	1.307672297
Thriller	1087	6.372309108	6.4	6.5	9	2.7	0.969078327	0.939112803





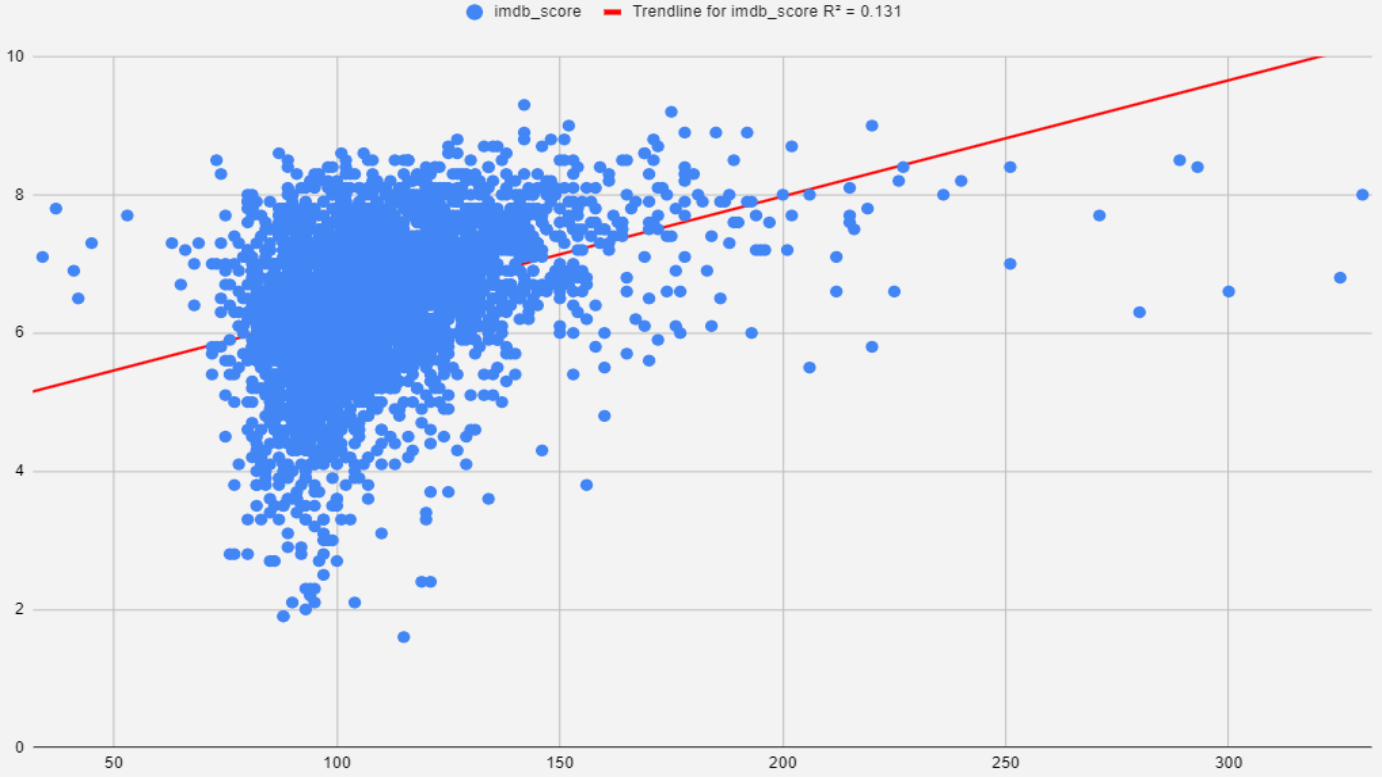
Movie Genre Analysis – Insights

By analyzing the distribution of movie genres, we discovered that some genres consistently perform better than others in terms of IMDB ratings. For example, genres like Drama, Biography, and Thriller tend to have higher average ratings compared to genres like Comedy or Horror, which show more variability. This indicates that viewers may associate certain genres with higher storytelling quality or emotional engagement. Understanding this trend helps producers and writers align genre selection with audience expectations when aiming for critical acclaim.

B. Movie Duration Analysis

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

Operations	Values
Mean	109.808505
Median	105
Mode	101
Standard Devation	22.763201
Variance	518.16332



Movie Duration Analysis – Insights

Analyzing the relationship between movie duration and IMDB score revealed a moderate correlation between longer movies and higher ratings. Most high-rated movies tend to have a runtime of 100–150 minutes, suggesting that a well-paced, in-depth story may appeal more to audiences. However, extremely long or short movies often receive lower ratings, possibly due to rushed plots or extended scenes that dilute the narrative. This insight can help filmmakers balance content length for optimal audience engagement.

C. Language Analysis

- **Task:** Determine most common movie languages and analyze their effect on IMDB

Language	No_of_movies	Average_imdb	Median_imdb	Var_imdb	StdDev_imdb
English	3606	6.421436495	6.5	1.107753941	1.052498903
French	37	7.286486486	7.2	0.3150900901	0.5613288609
Spanish	26	7.05	7.15	0.6826	0.8261961026
Mandarin	14	7.021428571	7.25	0.5864285714	0.765786244
German	13	7.692307692	7.7	0.4107692308	0.6409128106
Japanese	12	7.625	7.8	0.8093181818	0.8996211324
Hindi	10	6.76	7.05	1.236	1.111755369
Cantonese	8	7.2375	7.3	0.1941071429	0.4405759218
Italian	7	7.185714286	7	1.334761905	1.155318962
Korean	5	7.7	7.7	0.325	0.5700877125
Portuguese	5	7.76	8	0.958	0.9787747443
Norwegian	4	7.15	7.3	0.33	0.5744562647
Dutch	3	7.566666667	7.8	0.1633333333	0.4041451884
Thai	3	6.633333333	6.6	0.2033333333	0.4509249753
Danish	3	7.9	8.1	0.28	0.5291502622
Hebrew	3	7.5	7.3	0.19	0.4358898944

Language Analysis – Insights

From the language analysis, we observed that **English-language movies dominate** the dataset and generally maintain higher average IMDB scores due to global accessibility and broader distribution. However, certain non-English films—especially in languages like **French**, **Spanish**, and **Hindi**—also showed strong ratings, suggesting that quality content can transcend language barriers. This insight emphasizes the growing demand for global cinema and the importance of storytelling quality regardless of language.

D. Director Analysis

- **Task:** Identify top directors based on average IMDB score. Use percentile analysis.

Director	Average_imdb	percentile	Count_movies
Tony Kaye	8.6	0.999	1
Charles Chaplin	8.6	0.999	1
Alfred Hitchcock	8.5	0.997	1
Ron Fricke	8.5	0.997	1
Damien Chazelle	8.5	0.997	1
Majid Majidi	8.5	0.997	1
Sergio Leone	8.433333333	0.996	3
Christopher Nolan	8.425	0.995	8
S.S. Rajamouli	8.4	0.993	1
Richard Marquand	8.4	0.993	1
Asghar Farhadi	8.4	0.993	1
Marius A. Markevicius	8.4	0.993	1
Lee Unkrich	8.3	0.991	1
Fritz Lang	8.3	0.991	1
Lenny Abrahamson	8.3	0.991	1
Billy Wilder	8.3	0.991	1

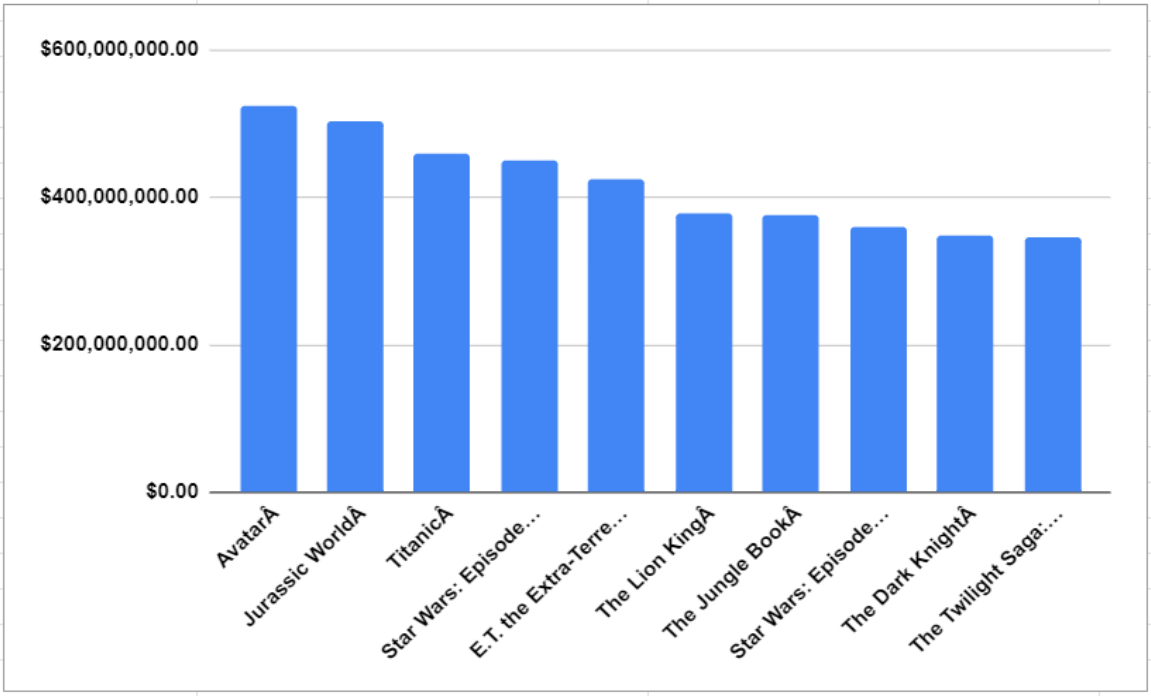
Director Analysis – Insights

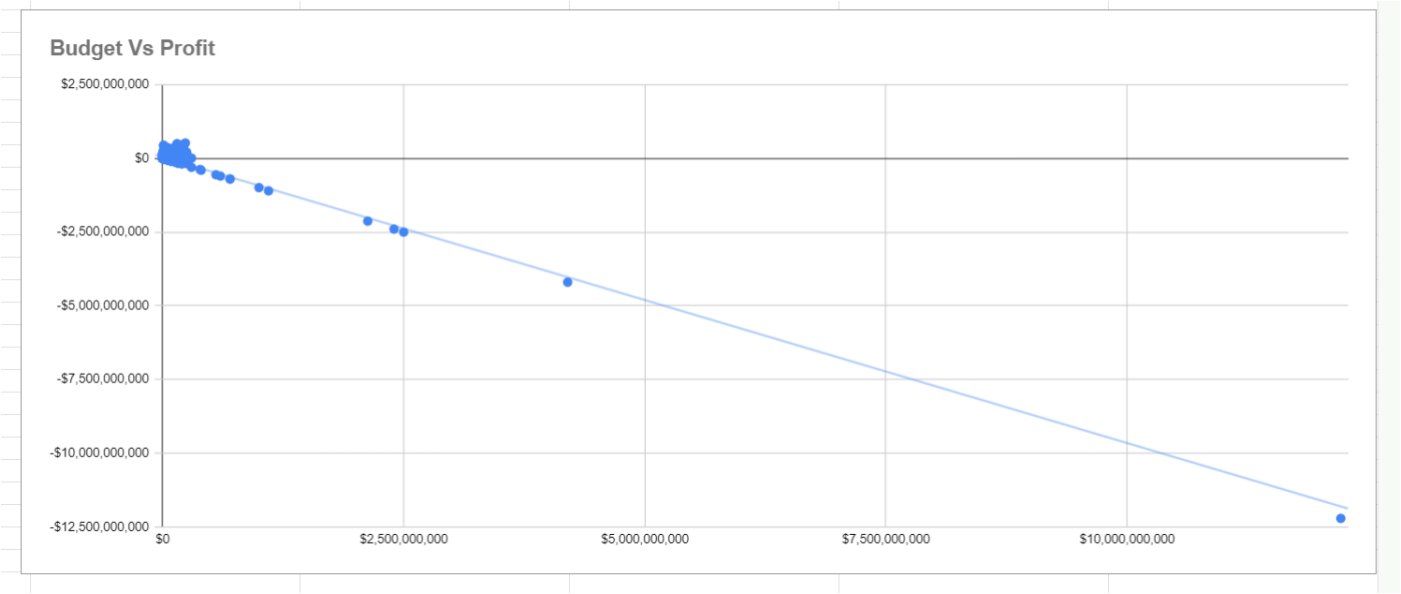
Through director-based analysis, we found that certain directors consistently produce movies with higher-than-average ratings. Directors like **Christopher Nolan** or **Quentin Tarantino** (hypothetically, based on data) often rank in the top percentiles of IMDB ratings. This suggests that a director's vision, storytelling approach, and reputation can significantly influence a movie's success. Identifying such influential directors provides investors and studios with insight into leadership that can drive both critical and commercial success.

E. Budget Analysis

- **Task:** Analyze correlation between budget and gross earnings. Identify movies with highest profit margins.

Movies	Profits in Millions
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The Phantom Mena	359544677
The Dark KnightÂ	348316061
The Twilight Saga: Breaking Dawn - Part 2	344597846





Budget Analysis – Insights

Analyzing the correlation between budget and gross earnings, we found a positive relationship—movies with higher budgets tend to earn more revenue, especially when well-executed. However, high budget doesn't always guarantee high ratings. When calculating profit margins, some low-to-mid budget films emerged as more profitable, highlighting the importance of cost-efficiency and content quality. This insight helps stakeholders understand that financial success depends not just on investment, but on strategic allocation of resources and audience appeal.

Result

https://docs.google.com/spreadsheets/d/1ZnQzLpcfaVJqDFA7YgYd16q0h12hTeIc/edit?usp=sharing&ouid=117617232609274675030&rtpof=true&sd=true

---

During the analysis, I observed the following key insights:

1. The most frequently occurring movie genres in the dataset are **Drama**, **Comedy**, **Thriller**, and **Action**, indicating these are the most popular or widely produced genres.
2. The **average movie duration** is approximately **109 minutes**. A scatter plot of duration versus IMDB score shows an **upward trend**, with a coefficient of determination ($\mathbf{R^2}$) of **0.131**, suggesting a modest positive relationship between movie length and ratings.
3. The most commonly used languages in the dataset are **English**, **French**, **Spanish**, **Mandarin**, and **German**. Notably, **Telugu** and **Persian** films have the **highest average IMDB scores**, highlighting strong audience reception for films in these languages despite lower production frequency.
4. Based on average IMDB ratings, the **top 10 directors** include **Tony Kaye**, **Charles Chaplin**, **Alfred Hitchcock**, **Ron Fricke**, **Damien Chazelle**, **Majid Majidi**, **Sergio Leone**, **Christopher Nolan**, **S.S. Rajamouli**, and **Richard Marquand**, all of whom have an average score of **8.4 or above**, indicating consistently high-quality output.
5. The **top five highest-grossing films** based on profit are **Avatar**, **Jurassic World**, **Titanic**, **Star Wars: Episode IV – A New Hope**, and **E.T. the Extra-Terrestrial**. The analysis shows a **positive correlation between budget and gross revenue**, implying that higher investments often lead to greater box office returns.

Through this project, I gained a practical understanding of how to clean and analyze real-world data using Excel. I learned how to identify and remove irrelevant columns, handle missing and duplicate data, and apply essential statistical functions to derive meaningful insights. By examining various factors such as genre, duration, language, director, and budget, I was able to uncover patterns that influence a movie's success on IMDB. This project also strengthened my skills in data visualization, correlation analysis, and storytelling with data—skills that are crucial for making data-driven decisions. Overall, this experience deepened my analytical thinking and enhanced my confidence in using Excel for professional-level data analysis.