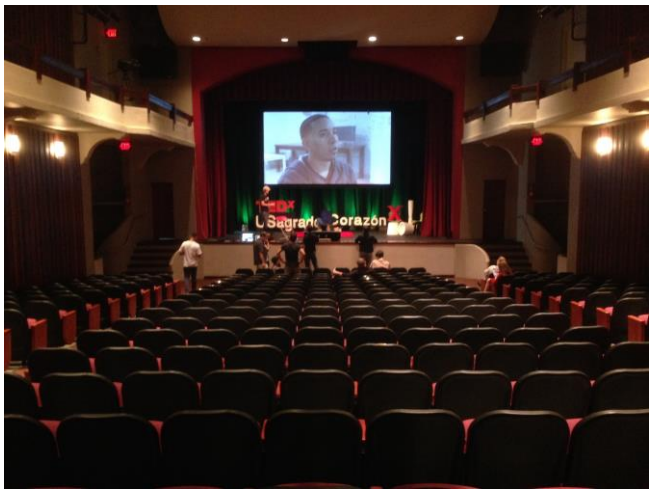


IMDB MOVIE ANALYSIS

PROJECT REPORT

This project analyzes IMDB movie data using excel to uncover trends in genre, duration, language, director performance, and profitability. Key insights include popular genres, ideal movie length, top-performing directors, and a weak correlation between budget and gross revenue. The analysis provides data-driven insights for decision-making in the film industry.

Deepthy A
Trainity Data Analytic Training

Project Description

The IMDB Movie Analysis project was designed to explore the elements that contribute to a movie's success, both critically and commercially. By examining a dataset of over 5,000 movies, the goal was to uncover how various factors - such as genre, duration, language, director, and budget - relate to a film's IMDB rating and financial performance. This analysis aims to draw insights that can inform future decisions in filmmaking, marketing, and production strategy.

Approach

Our approach began with cleaning and organizing the dataset to ensure accuracy and consistency. From there, we divided the analysis into five focused areas:

1. **Genre Analysis** - Evaluated how different genres perform on average in terms of IMDB scores.
2. **Movie Duration Analysis** - Explored how movie length affects viewer ratings, using descriptive statistics and visualizations.
3. **Language Analysis** - Assessed the frequency of languages used in movies and how language impacts IMDB ratings.
4. **Director Analysis** - Identified top-performing directors based on average scores and percentile rankings.
5. **Budget Analysis** - Investigated the financial side by analyzing the relationship between budgets, earnings, and profit margins.

Each section combined statistical functions to gain a clear, data-backed understanding of trends.

Tech Stack Used

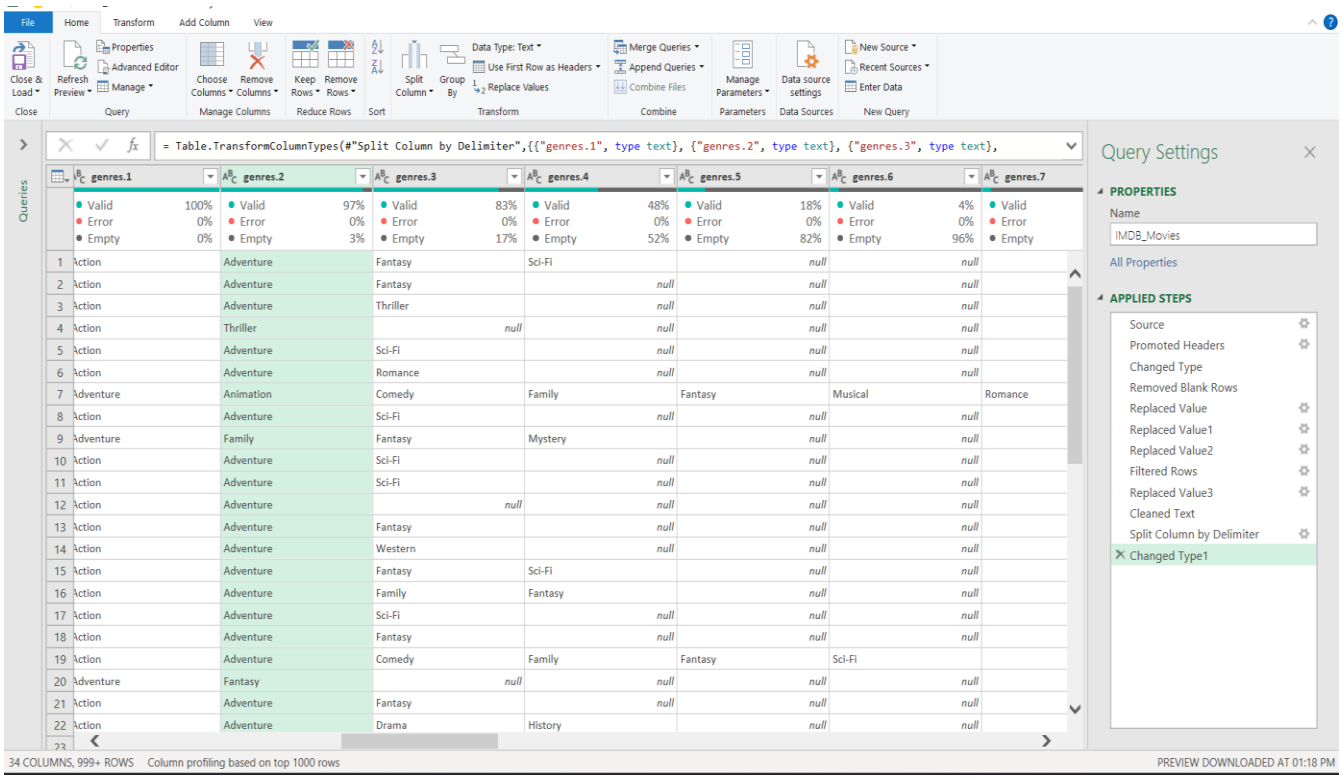
- **Microsoft Excel 365**
  Used extensively for data cleaning, statistical analysis (using formulas like AVERAGE, MEDIAN, STDEV, CORREL, etc.), pivot tables, and chart creation for visual insights.
- **Microsoft Word 365**
  Used for compiling, formatting, and documenting the report in a structured and professional format suitable for presentation and submission.

Data Cleaning and Preprocessing

The initial dataset contained over 5,000 movie records with details such as genre, duration, language, director, budget, gross earnings, and IMDB scores. To ensure high data quality and reliable insights, several cleaning steps were carried out using **Power Query** in **Microsoft Excel 365**:

- **Removed blank rows and null entries**: Rows with empty or missing critical fields were filtered out to maintain consistency.
- **Replaced inconsistent text entries**: Common issues such as inconsistent capitalization and extra spaces in genres and languages were resolved.
- **Split multi-genre fields**: Movies with multiple genres were split into separate columns for more granular analysis.
- **Changed data types**: Ensured numerical fields like "duration", "IMDB score", "budget", and "gross" were correctly formatted as numbers.
- **Applied transformations**: Used Excel formulas (e.g., TRIM, LOWER, IFERROR) and Power Query transformations to clean and structure the data efficiently.

These preprocessing steps allowed for smooth execution of subsequent analysis tasks and improved the accuracy of insights generated.



Movie Genre Analysis

Objective: To examine how different movie genres are associated with varying IMDB scores and assess which genres perform better overall.

Approach:

- Extracted genre information using Power Query by splitting separated values into distinct genre columns.
- Transformed the data to consolidate genre frequency and calculate **median**, **mode**, and **average IMDB score** per genre using **Pivot Tables** and **Excel's statistical functions**.

Genre	Count	Median	Mode
Action	1147	6.3	6.1
Adventure	921	6.6	6.7
Fantasy	609	6.4	6.7
Sci-Fi	615	6.4	6.7
Thriller	1401	6.4	6.1
Romance	1105	6.5	6.5
Animation	241	6.7	6.7
Comedy	1867	6.3	6.7
Family	545	6.4	6.7
Musical	132	6.7	7
Mystery	498	6.6	6.6
Western	97	6.8	6.5
Drama	2582	6.9	7.2
History	206	7.2	7.5
Sport	182	6.8	7.2
Crime	884	6.6	6.6
Horror	561	5.9	6.2
War	212	7.1	7.1
Biography	293	7.2	7
Music	214	6.6	6.5
Documentary	120	7.4	7.5
Game-Show	1	2.9	No Mode
Reality-TV	2	4.75	No Mode
News	3	7.4	No Mode
Short	5	6.5	No Mode
Film-Noir	6	7.65	No Mode

Row Labels	Count of imdb_score	Var of imdb_score	StdDev of imdb_score	Min of imdb_score	Average of imdb_score	Max of imdb_score
Action	1147	1.240020008	1.113561856	1.7	6.237314734	9
Adventure	921	1.276122646	1.129655985	1.9	6.439087948	8.9
Animation	241	1.30269917	1.141358476	1.7	6.578423237	8.6
Biography	293	0.52202908	0.722515799	4.5	7.150170648	8.9
Comedy	1867	1.1922221	1.091889234	1.7	6.194643814	9.5
Crime	884	1.050573118	1.024974691	2.4	6.564027149	9.3
Documentary	120	1.125595938	1.060941063	1.6	7.180833333	8.7
Drama	2582	0.90936991	0.953608887	2	6.767505809	9.3
Family	545	1.446381409	1.202655981	1.7	6.245504587	8.7
Fantasy	609	1.349252063	1.161573099	1.7	6.307553366	8.9
Film-Noir	6	0.186666667	0.43204938	7.1	7.633333333	8.2
History	206	0.78747336	0.887396957	2	7.088349515	8.9
Horror	561	1.266846893	1.125542933	2.2	5.853832442	8.7
Music	214	1.389659076	1.178838019	1.6	6.410280374	8.5
Musical	132	1.502384918	1.225718123	2.1	6.507575758	8.5
Mystery	498	1.188719506	1.09028414	2.2	6.48935743	8.6
News	3	0.263333333	0.513160144	7.1	7.533333333	8.1
Reality-TV	2	6.845	2.61629509	2.9	4.75	6.6
Romance	1105	0.993480507	0.996734923	2.1	6.450135747	8.6
Sci-Fi	615	1.458118588	1.207525813	1.9	6.277723577	8.8
Short	5	0.557	0.746324326	5.2	6.38	7.1
Sport	182	1.214272661	1.101940407	2	6.606043956	8.7
Thriller	1401	1.099543683	1.048591285	2.2	6.319057816	9
War	212	0.765911428	0.875163658	2.7	7.074056604	8.6
Western	97	1.086767612	1.042481468	3.8	6.689690722	8.9
Grand Total	14448	1.205277697	1.0978514	1.6	6.452249446	9.5

Key Findings:

- The most frequent genres were **Drama (2582 movies)**, **Comedy (1867)**, and **Action (1147)**.
- Genres such as **Comedy**, **Crime**, and **Drama** displayed higher median IMDB scores.
- Less common genres like **Film-Noir**, though fewer in number, had the highest median (7.65) and average (7.63) IMDB scores.
- Popular genres like **Action** and **Comedy** had large sample sizes but relatively lower median scores (6.3 and 6.4).

Insights:

- Niche or information-rich genres are rated more highly by audiences, possibly due to deeper storytelling or educational value.
- Commercial genres like **Comedy** and **Action** are more widespread but tend to cluster around average ratings.

Movie Duration Analysis

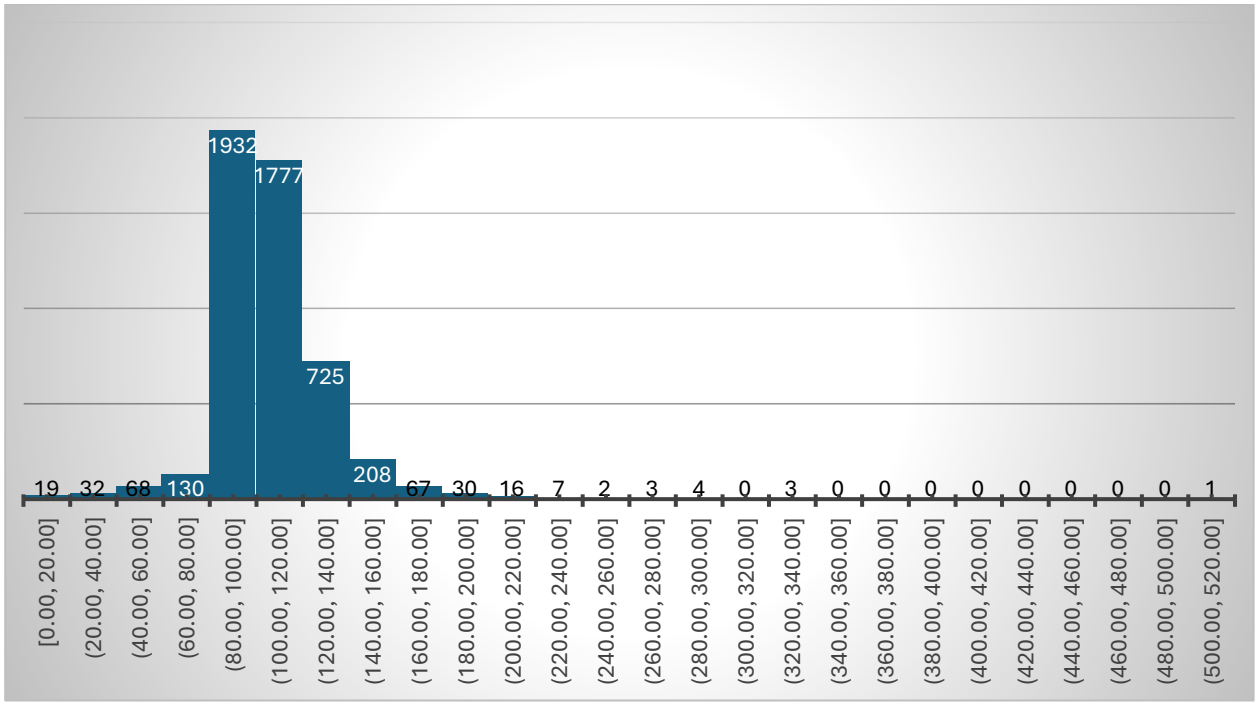
Objective:

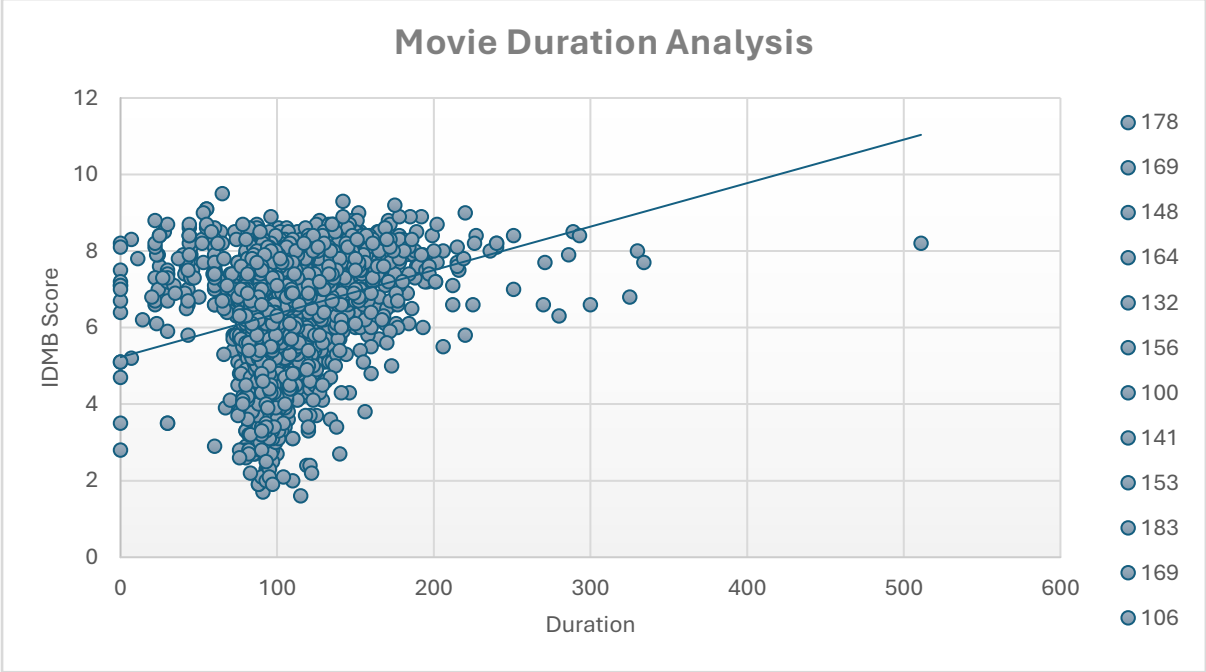
To explore the relationship between a movie's **duration** and its **IMDB score**, identifying any trends or outliers.

Approach:

- Used a **scatter plot with a trendline** to visualize the correlation between Duration and IMDB Score.
- Calculated **Mean**, **Median**, and **Standard Deviation** of durations for each genre to understand how movie lengths vary by category.

Genre	Mean	Median	Stdev
Action	110.8387	107	1.113562
Adventure	112.0456	107	1.129656
Fantasy	103.087	100	1.161573
Sci-Fi	107.187	104	1.207526
Thriller	107.935	106	1.048591
Romance	108.867	105	0.996735
Animatior	88.54357	90	1.141358
Comedy	98.91912	98	1.091889
Family	97.00183	95	1.202656
Musical	109.5076	103	1.225718
Mystery	106.51	105	1.090284
Western	120.1443	110	1.042481
Drama	112.6979	109	0.953609
History	136.5485	131	0.887397
Sport	110.2253	108.5	1.10194
Crime	108.336	105	1.024975
Horror	97.89483	96	1.125543
War	130.9245	124.5	0.875164
Biography	125.0239	121	0.722516
Music	108.3084	107	1.178838
Document	93.70833	92.5	1.060941
Game-Shc	60	60	no stdev
Reality-TV	51.5	51.5	2.616295
News	98.66667	105	0.51316
Short	26.4	34	0.746324
Film-Noir	99.5	93.5	0.432049





Interpretation:

- **Longer movies** (especially History, Biography, War, and Drama) generally correlate with **higher ratings**.
- **Short formats** like Game Shows, Reality-TV, and Shorts have **lower durations and lower average ratings**, suggesting audience preferences might lean toward longer, narrative-rich content.

- **Standard deviations are low**, showing that most genres have consistent duration patterns, with **Reality-TV** being an exception (highest variation).

Key Visual Insight:
The scatter plot reveals a **slight positive correlation** between movie duration and IMDB score. As the duration increases, IMDB scores also tend to rise marginally. However, the spread is wide, indicating that **duration alone is not a strong predictor** of a high rating.

Language Analysis

Objective:

To determine whether **movie language** influences **IMDB scores**, by analyzing average, median ratings and standard deviation per language.

Approach:

- Aggregated data for each language including **Mean**, **Median**, and **Standard Deviation** of IMDB ratings.
- Highlighted high-performing and low-performing languages.
- Created a **conditional formatting heatmap** for quick visual insights.

Language	Count	Mean	Median	Stdev
English	4692	6.400554	6.5	1.120447972
Japanese	18	7.394444	7.6	0.990823913
French	73	7.038356	7.2	0.726985812
Mandarin	25	6.776	7	1.061555463
Aboriginal	2	6.95	6.95	0.777817459
Spanish	40	6.9375	7.15	0.855056603
Filipino	1	6.7	6.7	
Hindi	28	6.632143	6.95	1.398955582
Russian	11	6.363636	6.5	1.383671007
Maya	1	7.8	7.8	
Kazakh	1	6	6	
Telugu	1	8.4	8.4	
Cantonese	11	6.954545	7.2	0.704788814
Icelandic	2	7.55	7.55	0.919238816
German	19	7.342105	7.6	0.954123093
Aramaic	1	7.1	7.1	
Italian	11	7.227273	7.3	1.244259546
Dutch	4	7.425	7.45	0.434932945
Dari	2	7.5	7.5	0.141421356
Hebrew	5	7.58	7.6	0.334664011
Chinese	3	5.666667	5.7	0.550757055
Mongolian	1	7.3	7.3	
Swedish	5	7.44	7.6	0.756967635
Korean	8	7.3875	7.5	0.825378701
Thai	3	6.633333	6.6	0.450924975
Polish	4	8.25	8.25	0.981495458
Bosnian	1	4.3	4.3	
None	2	7.95	7.95	0.777817459
Hungarian	1	7.1	7.1	
Portuguese	8	7.4875	7.7	0.883883476

Unknown	6	6.8	6.95	1.331164903
Danish	5	7.5	8.1	1.077032961
Arabic	5	7.38	7.4	0.884307639
Norwegian	4	7.15	7.3	0.574456265
Czech	1	7.4	7.4	
Kannada	1	7.1	7.1	
Zulu	2	7.1	7.1	0.282842712
Panjabi	1	6.6	6.6	
Tamil	1	5.1	5.1	
Dzongkha	1	7.5	7.5	
Vietnamese	1	7.4	7.4	
Indonesian	2	7.9	7.9	0.424264069
Urdu	1	7	7	
Romanian	2	7.2	7.2	0.989949494
Persian	4	7.575	7.95	1.203813385
Slovenian	1	6.4	6.4	
Greek	1	7.3	7.3	
Swahili	1	7.4	7.4	

Key Insights:

1. **Highest Rated Languages (Mean Rating > 7.5)**
    - **Telugu (8.4)** - Highest rated, though based on a single entry.
    - **Polish (8.25)** - Consistently high rating, stronger due to multiple entries.
    - **Maya (7.8)**, **Indonesian (7.9)**, and **Hebrew(7.58)** also show strong average ratings with low standard deviations, indicating consistency.
2. **Popular Languages with High Volume:**
    - **English** (4,692 movies): Mean score **6.4**, Median **6.5**, with a wider spread (Stdev: 1.12).
        - Indicates **broad diversity** in English-language films ranging from low to high-rated.
    - **French** (73): High mean rating (**7.03**) and low deviation (**0.72**) - shows reliable performance.
    - **Japanese** and **German** are also strong performers with means above 7.3.
3. **Languages with Lower Scores:**
    - **Chinese** (Mean: 5.67), **Bosnian** (4.3), and **Tamil** (5.1) are among the **lowest rated**.
    - These may represent niche or limited-audience productions, or limited data points.

4. **Languages with Consistent Ratings (Low Std Dev):**
    - **Dari (0.14)**, **Slovenian (N/A)**, **Dutch (0.43)**, and **Indonesian (0.42)** show **tight clustering**, suggesting **strong consistency** in ratings.

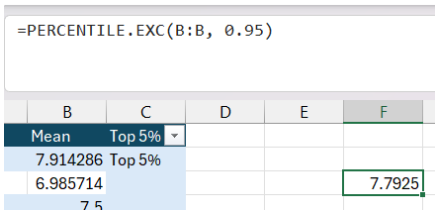
Director Analysis

Objective:

To identify the top-performing directors based on average IMDb movie ratings and understand the characteristics of high-rated direction.

Approach:

- Calculated the 95th percentile threshold using the formula:
  =PERCENTILE.EXC(B:B, 0.95)



- Directors with a mean rating equal to or above **7.79** were classified as **Top 5%**.
- Tagged and extracted data for directors meeting the criteria.

Director	Mean	Top 5%
James Cameron	7.914286	Top 5%
Christopher Nolan	8.425	Top 5%
Nathan Greno	7.8	Top 5%
Joss Whedon	7.925	Top 5%
Lee Unkrich	8.3	Top 5%
Pete Docter	8.233333	Top 5%
Don Hall	7.9	Top 5%
Rich Moore	7.8	Top 5%
Hideaki Anno	8.2	Top 5%
Alfonso Cuarón	7.8	Top 5%
Quentin Tarantino	8.2	Top 5%
Jacques Perrin	7.9	Top 5%
Frank Darabont	7.975	Top 5%
Stanley Kubrick	8	Top 5%
Tim Miller	8.1	Top 5%
Milos Forman	8.133333	Top 5%
Deepa Mehta	7.8	Top 5%
Andrei Tarkovsky	8.1	Top 5%
Denis Villeneuve	7.966667	Top 5%
S.S. Rajamouli	8.4	Top 5%
Moustapha Akkad	8.4	Top 5%

Tony Kaye	8.033333	Top 5%
Hayao Miyazaki	8.225	Top 5%
Richard Marquand	8.4	Top 5%
Sergio Leone	8.475	Top 5%
David Lean	8	Top 5%
Bernardo Bertolucci	7.95	Top 5%
Giuseppe Tornatore	7.8	Top 5%
Christian Carion	7.8	Top 5%
Tom McCarthy	7.9	Top 5%
James Schamus	7.8	Top 5%
Terry George	8.1	Top 5%
George Cukor	7.9	Top 5%
Michel Hazanavicius	8	Top 5%
Morten Tyldum	7.85	Top 5%
Catherine Owens	8.4	Top 5%
Fritz Lang	8.3	Top 5%
Andrey Zvyagintsev	8	Top 5%
John Blanchard	9.5	Top 5%
Lenny Abrahamson	8.3	Top 5%
Stephen Chbosky	8	Top 5%
John Cromwell	7.8	Top 5%
Je-kyu Kang	8.1	Top 5%
Stéphane Aubier	7.9	Top 5%
Josh Boone	7.8	Top 5%
Akira Kurosawa	8.1	Top 5%
Ken Annakin	7.8	Top 5%
Stanley Kramer	7.95	Top 5%
Dan Gilroy	7.9	Top 5%
Jonathan Dayton	7.9	Top 5%
Justin Tipping	7.8	Top 5%
Vincent Paronnaud	8	Top 5%
George Roy Hill	8.2	Top 5%
Robert Stevenson	7.8	Top 5%
Christophe Barratier	7.9	Top 5%
Rakeysh Omprakash Mehra	8.4	Top 5%
Mike Mayhall	8.6	Top 5%
Shona Auerbach	7.8	Top 5%
Mike van Diem	7.8	Top 5%
Lukas Moodysson	7.9	Top 5%
Raja Menon	8.5	Top 5%
Billy Wilder	7.975	Top 5%
John Sturges	8.3	Top 5%
Ron Fricke	8.5	Top 5%
Howard Hughes	7.8	Top 5%
Jim Abrahams	7.8	Top 5%
Richard Brooks	7.8	Top 5%
Jay Oliva	8.4	Top 5%
Charles Ferguson	7.866667	Top 5%

Damien Chazelle	8.5	Top 5%
Frank Capra	8.06	Top 5%
Howard Hawks	7.8	Top 5%
Mel Stuart	7.8	Top 5%
Rajkumar Hirani	8.2	Top 5%
Stacy Peralta	7.8	Top 5%
Stanley Donen	8.3	Top 5%
William Wyler	8.1	Top 5%
Robert Mulligan	8.4	Top 5%
Robert Rossen	8	Top 5%
Sylvain Chomet	7.8	Top 5%
Elia Kazan	7.866667	Top 5%
Ari Folman	8	Top 5%
David Sington	8.1	Top 5%
Ralph Ziman	7.8	Top 5%
Jehane Noujaim	8.1	Top 5%
Mitchell Altieri	8.7	Top 5%
Ritesh Batra	7.8	Top 5%
Fabián Bielinsky	7.9	Top 5%
Charles Chaplin	8.6	Top 5%
Frank Lotito	8.2	Top 5%
Sadyk Sher-Niyaz	8.7	Top 5%
Anna Muylaert	7.9	Top 5%
Robert Kenner	7.9	Top 5%
Joshua Oppenheimer	8.2	Top 5%
Cristian Mungiu	7.9	Top 5%
Asghar Farhadi	8.4	Top 5%
Georg Wilhelm Pabst	8	Top 5%
Mark Sandrich	7.8	Top 5%
Michael Wadleigh	8.1	Top 5%
Lance McDaniel	8	Top 5%
Marius A. Markevicius	8.4	Top 5%
Justin Paul Miller	8.3	Top 5%
Joe Kenemore	8.2	Top 5%
Ingmar Bergman	8.2	Top 5%
D.W. Griffith	8	Top 5%
Henry Alex Rubin	7.8	Top 5%
Lauren Lazin	8	Top 5%
Michael Roemer	8.1	Top 5%
Sam Martin	7.8	Top 5%
Kristin Rizzo	8	Top 5%
Carl Theodor Dreyer	8.1	Top 5%
Sharon Greytak	8.1	Top 5%
Majid Majidi	8.5	Top 5%
Cary Bell	8.7	Top 5%
Bill Melendez	8.4	Top 5%
Amal Al-Agroobi	8.2	Top 5%
Sut Jhally	8.3	Top 5%

Dave Carroll	7.9	Top 5%
Anthony Vallone	7.8	Top 5%

Key Insights:

- The **Top 5% threshold** was determined to be **7.79**.
- **John Blanchard** holds the highest average rating of **9.5**.
- Some other directors are:

Director	Mean
John Blanchard	9.5
Mike Mayhall	8.6
Raja Menon	8.5
Ron Fricke	8.5
Damien Chazelle	8.5
Mitchell Altieri	8.7
Charles Chaplin	8.6
Sadyk Sher-Niyaz	8.7
Majid Majidi	8.5
Cary Bell	8.7

- A total of **118 directors** fell into the Top 5% category.
- Many top directors have strong critical acclaim, cross-genre consistency, or a history of award-winning work.

Budget Analysis

Objective:

To explore the relationship between a movie's budget and its gross revenue and identify the most and least profitable titles.

Approach:

- Calculated the **profit margin** using the formula:

  = gross - budget

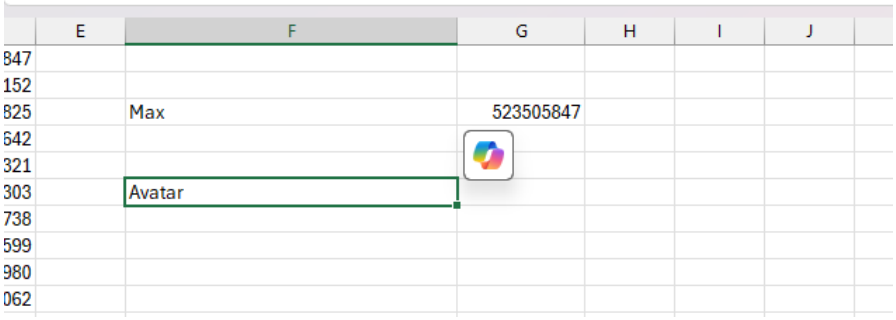
- Used the CORREL() function to measure the linear relationship between **budget** and **gross**.

```
=CORREL(IMDB_Movies[budget],IMDB_Movies[gross])
```

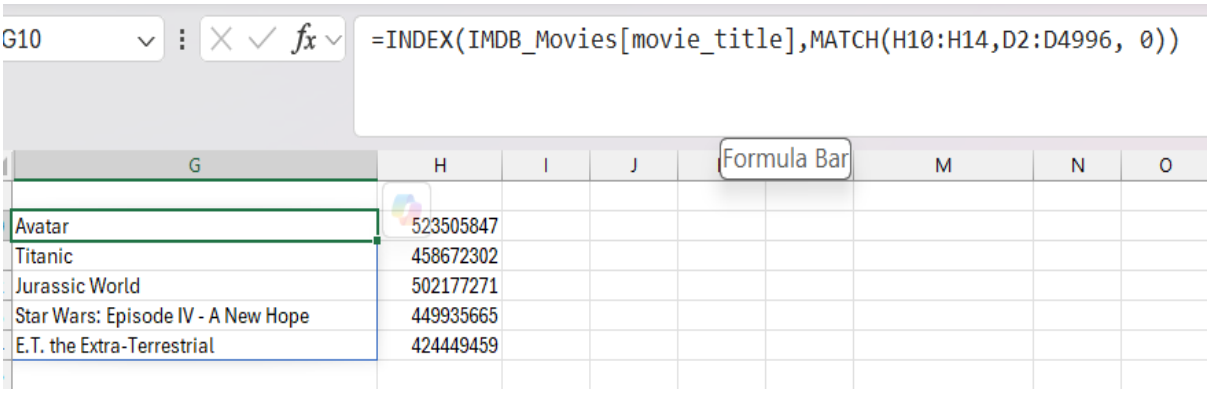
0.102166

- Identified the **highest** profit margin movie.

```
=INDEX(IMDB_Movies[movie_title],MATCH(MAX(D:D), D:D, 0))
```



- Found top 5 most profitable movies.



Key Insights:

- The **correlation coefficient** between budget and gross is **0.10**, suggesting a **very weak positive correlation**.
- This indicates that a **higher budget does not strongly guarantee higher revenue**.
- The **most profitable movie** is **"Avatar".**
- Top 5 profitable movies are:

Avatar
Titanic
Jurassic World
Star Wars: Episode IV - A New Hope
E.T. the Extra-Terrestrial

- Profitability varies widely, indicating that **factors beyond budget** (like content quality, marketing, or star power) play a major role in box office success.

Conclusion

This analysis of IMDB movie data provided valuable insights into the elements that contribute to a movie's success. We identified the most popular genres, ideal movie durations, top-performing languages and directors, and examined the financial performance of films. While budget and gross revenue showed only a weak correlation, the study highlights that factors like genre and director can significantly influence a movie's reception. These findings can support better planning and decision-making for filmmakers and studios aiming to optimize both audience engagement and profitability.

Drive Link:
https://docs.google.com/spreadsheets/d/1gsIZabkwBWDtmNqlN4IaboWrdxkfe2JT/edit?usp=sharing&ouid=112959782025131466050&rtpof=true&sd=true