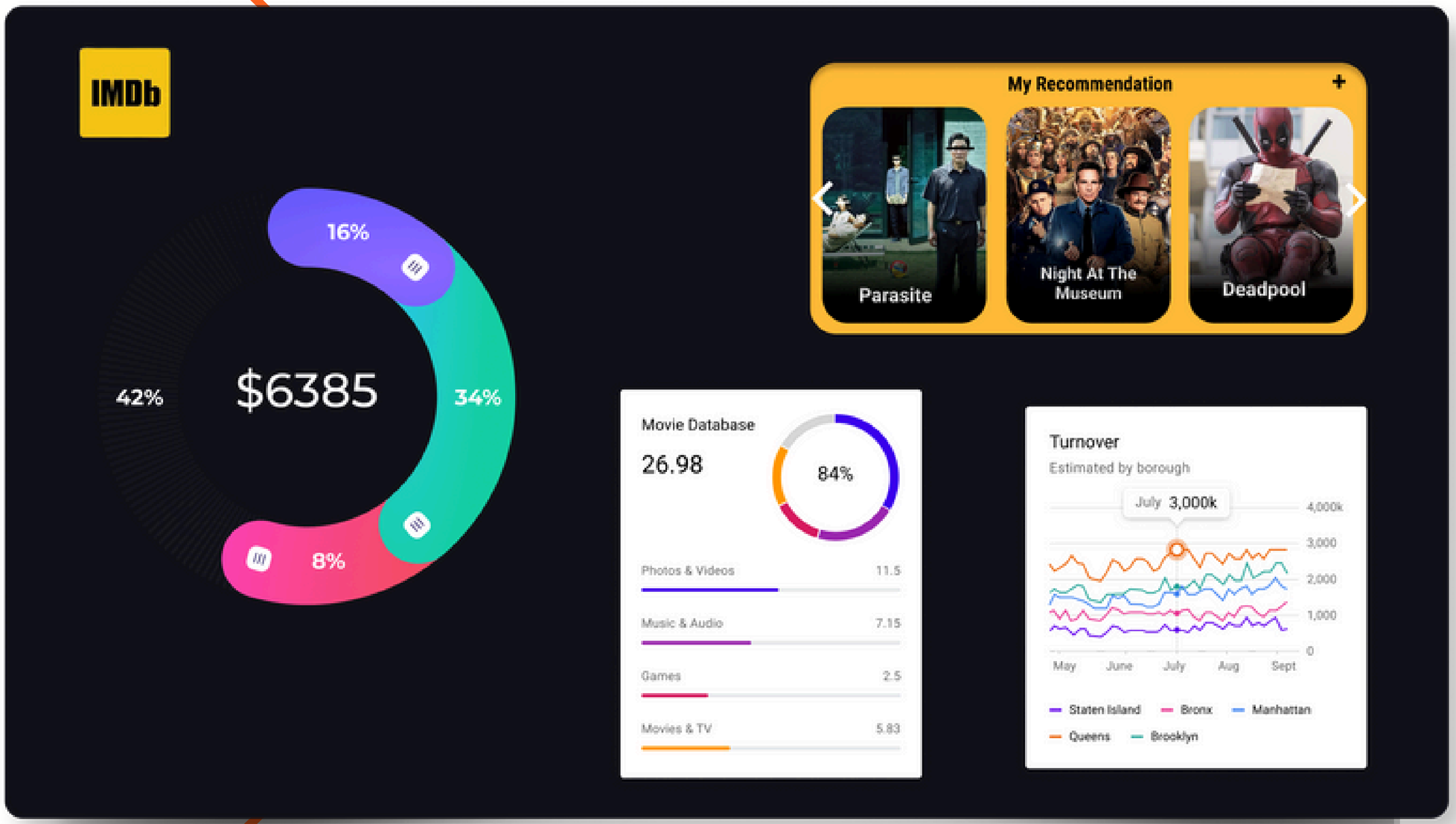
IMDB MOVIE ANALYSIS



by: Anirudh Chaudhary

PROJECT DESCRİPTİON:

This project focuses on analyzing IMDB movie data to uncover key trends and insights that influence movie success. By leveraging a comprehensive dataset, we aim to identify the factors that contribute to higher ratings, greater profitability, and audience engagement. The analysis spans critical elements such as movie genres, duration, language, budget, and director performance.
Scope of the Project:

- **Data Source:** IMDB database with detailed movie information (e.g., ratings, genres, budgets, etc.).
- **Timeframe:** Movies released across the past two decades to ensure a robust analysis.
- **Key Metrics:** IMDB ratings, gross earnings, profit margins, and audience reach.

Objective:
To provide actionable insights for stakeholders in the entertainment industry, including producers, directors, and investors, enabling them to make data-driven decisions that maximize both creative and financial outcomes.

Impact:
The findings of this project will serve as a strategic guide for optimizing movie production, marketing, and distribution strategies to align with market demands and audience preferences.

APPROACH:

To systematically analyze the IMDB movie dataset, the following structured approach was adopted, ensuring a balance between data accuracy, statistical depth, and actionable insights:

1. Data Understanding:

- Explored the dataset to understand its structure, attributes, and the relationships between variables.
- **Key Attributes:** IMDB ratings, genres, directors, languages, budget, gross earnings, and movie duration.
- Identified potential problem areas such as missing data, outliers, and inconsistencies.

2. Data Cleaning and Preparation:

- **Missing Values:** Addressed missing data by imputation (e.g., mean for numerical data, mode for categorical data) or removal when values were insignificant.
- **Duplicates:** Removed duplicate entries to avoid skewed results.
- **Data Transformation:** Converted columns like genres and languages into an analyzable format using text splitting and one-hot encoding.
- **Feature Engineering:** Created new metrics like profit margin (gross earnings - budget) and genre combinations.

3. Data Analysis:

- **Movie Genre Analysis:** Studied the frequency and distribution of genres. Computed statistical measures (mean, median, variance, etc.) of IMDB ratings across genres.
- **Duration Analysis:** Analyzed the impact of movie durations on ratings using scatter plots and trendlines.
- **Language Analysis:** Examined how movie languages influenced IMDB ratings through statistical comparisons.
- **Director Analysis:** Identified top-performing directors based on percentile rankings and average ratings.
- **Budget Analysis:** Assessed correlations between budgets and financial success using correlation coefficients and profit margin analysis.

TECH-STACK USED

Excel 2022 (with 365) was chosen for its versatility and ability to handle large datasets with ease. The following features were extensively used:

- **Data Cleaning:**
  - **Text-to-columns** for splitting genre and language data.
  - **Filtering and sorting** to identify and handle duplicates or missing values.

- **Descriptive Statistics:**
  - Used functions like **AVERAGE, MEDIAN, MODE, MAX, MIN, VAR, and STDEV** to compute statistical measures for various attributes (e.g., ratings, budgets, durations).

- **Data Visualization:**
  - Created **scatter plots, histograms, and bar charts** to visually represent relationships between variables.
  - Used **trendlines** to analyze correlations and patterns.

- **Advanced Functions:**
  - **COUNTIF** for genre and language frequency distribution.
  - **CORREL** for measuring the correlation between budgets and gross earnings.
  - **PERCENTILE** for ranking directors based on average ratings.

- **Pivot Tables:**
  - Generated **summary statistics** for genres, directors, and languages to uncover trends and insights.

DATA-CLEANING

Data Cleaning and Preparation:

- **Missing Values:** Addressed missing data by imputation (e.g., mean for numerical data, mode for categorical data) or removal when values were insignificant.
- **Duplicates:** Removed duplicate entries to avoid skewed results.
- **Data Transformation:** Converted columns like genres and languages into an analyzable format using text splitting and one-hot encoding.

DATA-CLEANING

2 Highlighted the blank cells, with cell-highlight (yellow)

color	director_na	num_critic_for_rev	duratio	director_facebook	actor_3_facebook	actor_2_na	actor_1_facebook_l	gross	genres
Color	James Camerc	723	178	0	855	Joel David Mo	1000	760505847	Action|Adventure|Fantasy|Sc
Color	Gore Verbinsk	302	169	563	1000	Orlando Bloo	40000	309404152	Action|Adventure|Fantasy
Color	Sam Mendes	602	148	0	161	Rory Kinnear	11000	200074175	Action|Adventure|Thriller
Color	Christopher N	813	164	22000	23000	Christian Bale	27000	448130642	Action|Thriller
	Doug Walker			131		Rob Walker	131		Documentary
Color	Andrew Stantc	462	132	475	530	Samantha Mc	640	73058679	Action|Adventure|Sci-Fi
Color	Sam Raimi	392	156	0	4000	James Franco	24000	336530303	Action|Adventure|Romance
Color	Nathan Greno	324	100	15	284	Donna Murph	799	200807262	Adventure|Animation|Comed
Color	Joss Whedon	635	141	0	19000	Robert Downe	26000	458991599	Action|Adventure|Sci-Fi
Color	David Yates	375	153	282	10000	Daniel Radclif	25000	301956980	Adventure|Family|Fantasy|My
Color	Zack Snyder	673	183	0	2000	Lauren Cohar	15000	330249062	Action|Adventure|Sci-Fi
Color	Bryan Singer	434	169	0	903	Marlon Brand	18000	200069408	Action|Adventure|Sci-Fi
Color	Marc Forster	403	106	395	393	Mathieu Amal	451	168368427	Action|Adventure

Cleaning the “Offered-Salary” Column, Identifying Blank Cells/Duplicate Values

DATA-CLEANING

2

Highlighted the blank cells, with cell-highlight (yellow) and deleted the duplicated values

5036	Color	Neill Dela Llana
5037	Color	Robert Rodriguez
5038	Color	Anthony Vallone
5039	Color	Edward Burns
5040	Color	Scott Smith
5041	Color	
5042	Color	Benjamin Roberds
5043	Color	Daniel Hsia
5044	Color	Jon Gunn
5045		

4992	Color	Robert Rodrig
4993	Color	Anthony Vallone
4994	Color	Edward Burns
4995	Color	Scott Smith
4996	Color	
4997	Color	Benjamin Rob
4998	Color	Daniel Hsia
4999	Color	Jon Gunn
5000		

MOVİE GENRE ANALYSİS

1) Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics of the IMDB scores.

1

2

3

4

5

6

7

8

9

1 **Data Preparation:** Split Multiple Genres: Use the Text-to-Columns feature or Power Query to separate movies with multiple genres into individual rows for accurate analysis.

2 **Split Multiple Genres:** Use the Text-to-Columns feature or Power Query to separate movies with multiple genres into individual rows for accurate analysis.

3 **Genre Distribution Analysis:**

- **Count Genre Frequency:** Use the COUNTIF function to calculate the number of movies for each genre.
- **Visualize Results:** Create a bar chart to represent the distribution of genres.

MOVİE GENRE ANALYSİS

4

Calculate Descriptive Statistics for IMDB Scores:

Mean: =AVERAGE(range) **Median:** =MEDIAN(range) **Mode**: =MODE.SNGL(range) **Range:** =MAX(range) - MIN(range) Variance: =VAR.P(range) **Standard Deviation:** =STDEV.P(range)

MOVİE GENRE ANALYSİS

MEAN	MEDIAN	MODE	RANGE	VARIANCE	STD. DEVIATION
280	7.65	7.5	2570	405070.24	636.4512864

Row Labels	Sum of TOTAL-COUNTS
Action	14.17%
Comedy	**23.08%**
Drama	**31.87%**
Romance	13.60%
Thriller	17.27%
Grand Total	**100.00%**

MOVİE DURATİON ANALYSİS:

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

1 Calculate Descriptive Statistics for Durations:

- **Mean Duration:** =AVERAGE(range)
- **Median Duration:** =MEDIAN(range)
- **Standard Deviation:** =STDEV.P(range)

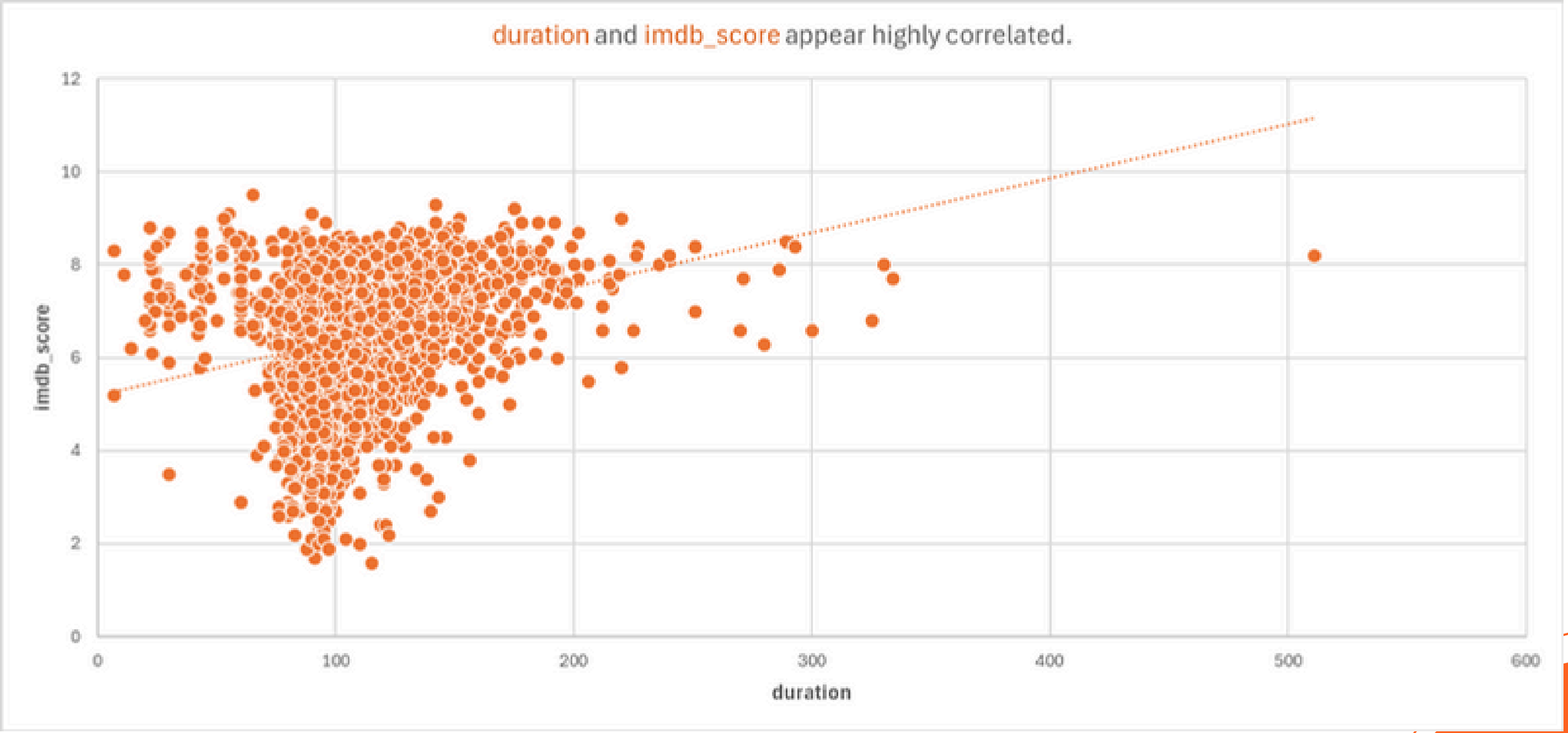
2 Analyze Relationship with IMDB Score:

- **Create a Scatter Plot with:**
  - X-Axis: Movie Duration
  - Y-Axis: IMDB Score

3 Add a Trendline to assess the relationship:

MOVİE DURATİON ANALYSİS:

MEAN	MEDIAN	Std. Deviation
107.21	103.00	25.25



MOVİE DURATİON ANALYSİS:

1) Task: Determine the most common languages used in movies and analyze their impact on the IMDB score using descriptive statistics.

1 Determine Language Distribution:

- Use COUNTIF to calculate the number of movies for each language:
- =COUNTIF(range, "Language")

2 Calculate Descriptive Statistics for IMDB Scores by Language:

- **Mean IMDB Score:** =AVERAGE(range) (filter by language).
- **Median IMDB Score:** =MEDIAN(range) (filter by language).
- **Standard Deviation:** =STDEV.P(range) (filter by language).

MOVİE DURATİON ANALYSİS:

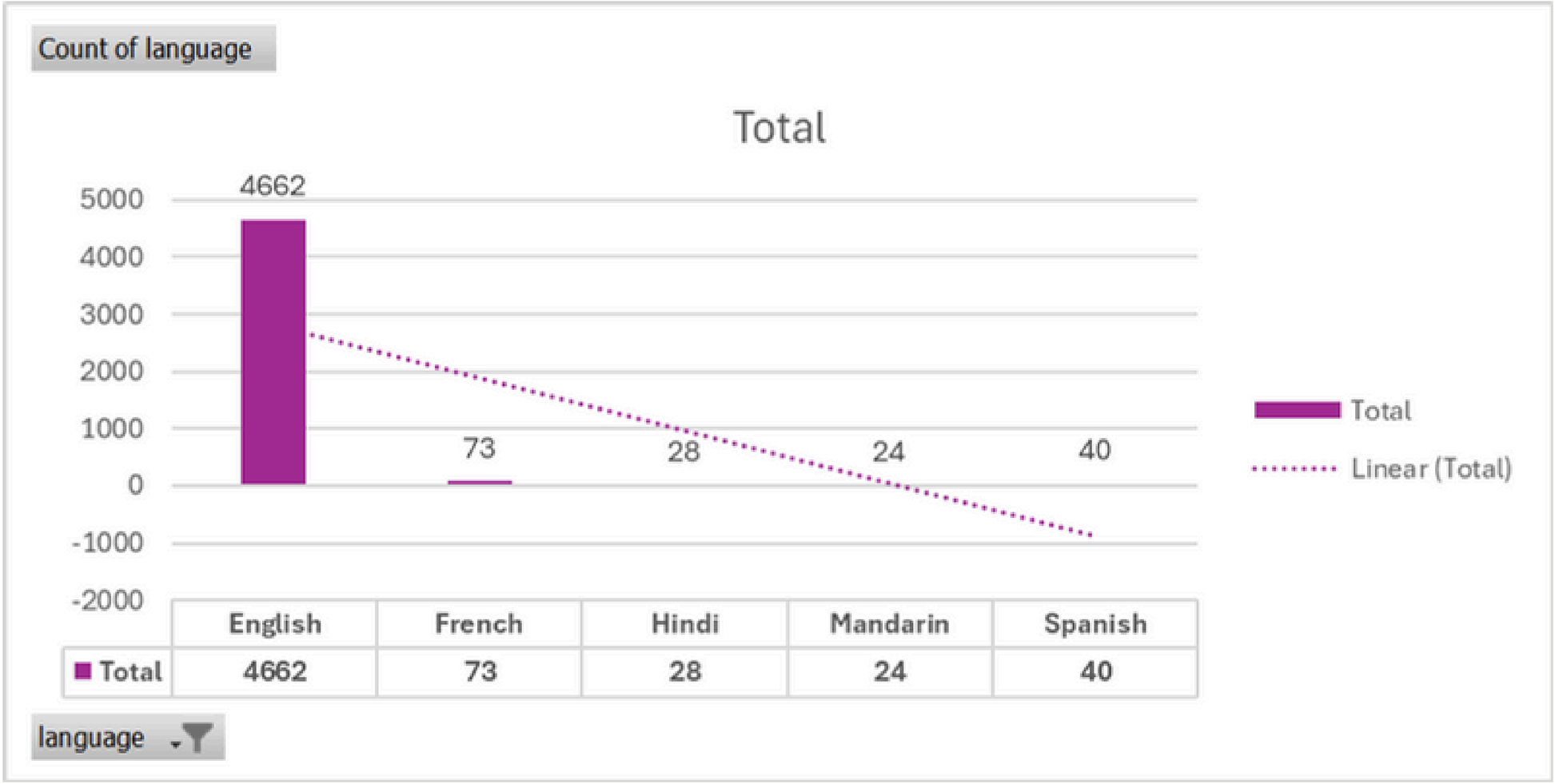
3

MEAN	MEDIAN	Std. Deviation
6.95	6.6	1.124107297
7.38	6.6	1.124030254
7.10	6.6	1.124104057
4.30	6.6	1.124205093
6.95	6.6	1.123939765
5.67	6.6	1.124013527
7.40	6.6	1.124123844
7.50	6.6	1.124231342
7.50	6.6	1.124179103
7.43	6.6	1.124191537
7.50	6.6	1.124203935
6.40	6.6	1.124297734
6.70	6.6	1.124400228

TOP 5 Languages

Row Labels	Count of language
English	4662
French	73
Hindi	28
Mandarin	24
Spanish	40
Grand Total	**4827**

MOVIE DURATION ANALYSIS:



MOVİE DURATİON ANALYSİS:

1) Task D: : Identify the top directors based on their average IMDB score and analyze their contribution to the success of movies using percentile calculations.

1

Calculate Average IMDB Score for Each Director:

- **Use a Pivot Table:**
- **Place 'Director'** in the Rows section.
- **Place 'IMDB Score' i**n the Values section, set to Average.

2

Alternatively, use the AVERAGEIF function:

=AVERAGEIF(range, "Director Name", IMDB_Score_Range)

MOVİE DURATİON ANALYSİS:

3

Identify Top Directors Using Percentiles:

- Calculate the Percentile of each director's average IMDB score

4

Visualize the Distribution:

- Create a Histogram / Pie Chart showing the distribution of average IMDB scores among directors.
- Highlight top-performing directors using a different color.

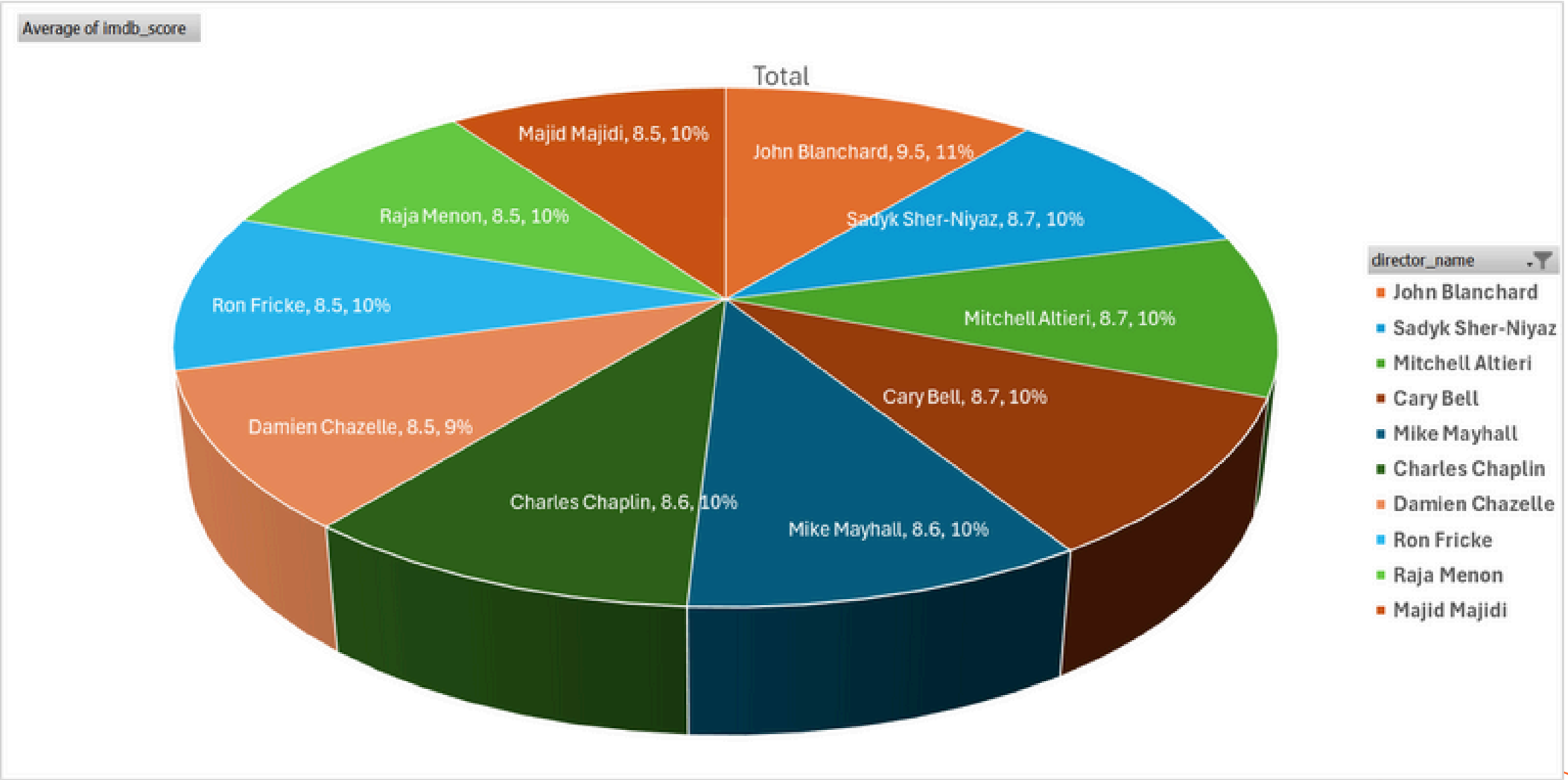
MOVIE DURATION ANALYSIS:

Row Labels	Average of imdb_score	90th Percentile	Top Directors
A. Raven Cruz	1.9	7.5	262
Ã‰mile Gaudreault	6.7		
Ã‰ric Tessier	6.6		
Ã‰tienne Faure	4.3		
Ãlex de la Iglesia	6.1		
Aaron Hann	6		
Aaron Schneider	7.1		
Aaron Seltzer	2.7		
Abel Ferrara	6.6		
Adam Brooks	7.2		
Adam Carolla	6.1		
Adam Goldberg	5.4		
Adam Green	5.7		

Top 10 Directors

Row Labels	Average of imdb_score
John Blanchard	9.5
Sadyk Sher-Niyaz	8.7
Mitchell Altieri	8.7
Cary Bell	8.7
Mike Mayhall	8.6
Charles Chaplin	8.6
Damien Chazelle	8.5
Ron Fricke	8.5
Raja Menon	8.5
Majid Majidi	8.5

MOVİE DURATİON ANALYSİS:



MOVİE DURATİON ANALYSİS:

1

Calculate the Correlation Between Budget and Gross Earnings:

- Use the CORREL function to calculate the correlation coefficient between movie budgets and gross earnings:
- **=CORREL(Budget_Range, Gross_Earnings_Range)**

2

Calculate Profit Margin:

- Calculate the profit margin (gross earnings - budget) for each movie:
- Profit Margin = Gross Earnings - Budget

MOVİE DURATİON ANALYSİS:

3 Identify Movies with the Highest Profit Margin:

- Use the MAX function to find the highest profit margin:
- **=MAX(Profit_Margin_Range)**

4 Visualize the Data:

- Create a Scatter Plot to visualize the relationship between budget and gross earnings.
- Add a Trendline to assess the strength and direction of the correlation.

MOVİE DURATİON ANALYSİS:

Movie_Name	gross_earning	budget	Profit
The HostÂ	2201412	12215500000	-12213298588.00
Lady VengeanceÂ	211667	4200000000	-4199788333.00
FatelessÂ	195888	2500000000	-2499804112.00
Princess MononokeÂ	2298191	2400000000	-2397701809.00
SteamboyÂ	410388	2127519898	-2127109510.00
AkiraÂ	439162	1100000000	-1099560838.00

Correlation
0.101033478

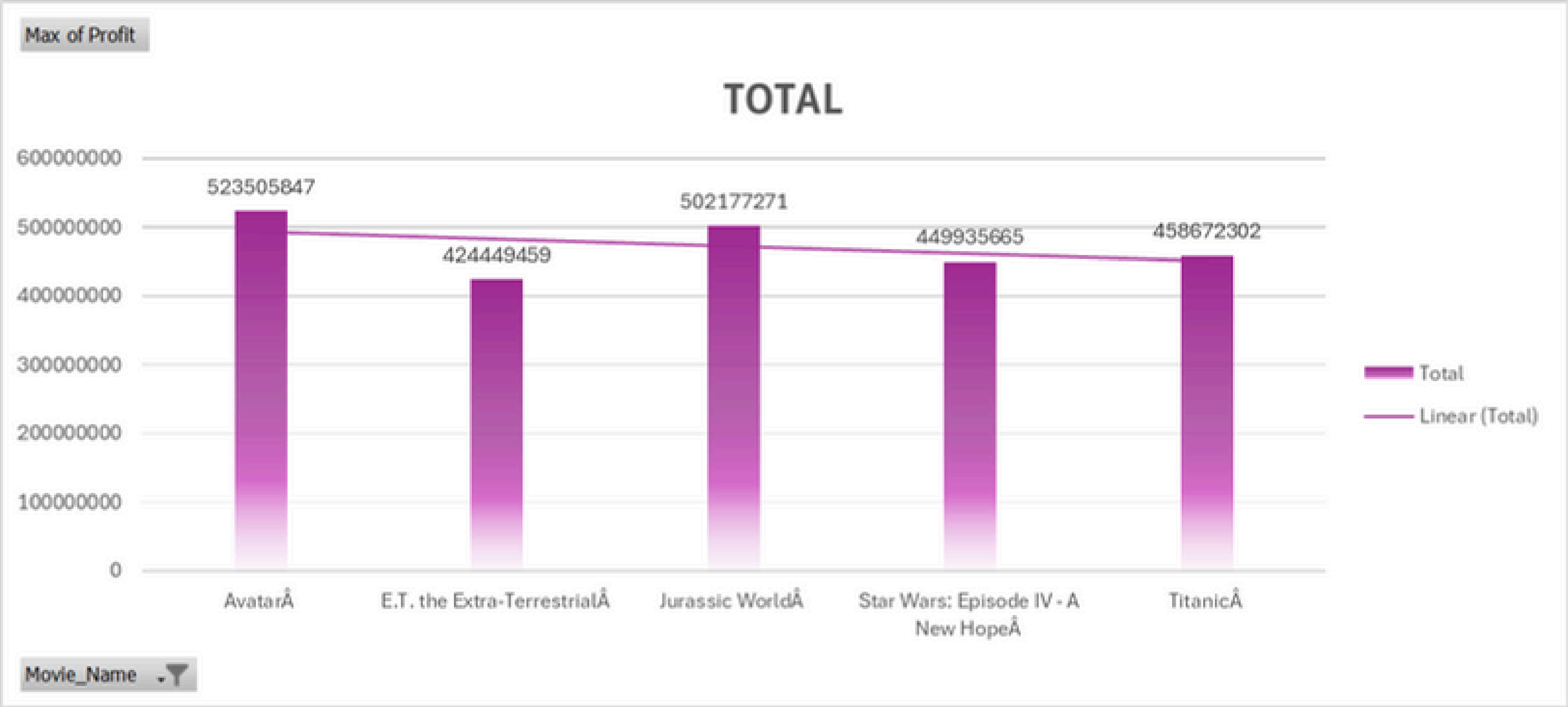
Highest Profit Margin
523505847

MOVİE DURATİON ANALYSİS:

TOP 5 with Highest Earning	
Row Labels	Max of Profit
AvatarÂ	523505847
E.T. the Extra-TerrestrialÂ	424449459
Jurassic WorldÂ	502177271
Star Wars: Episode IV - A New HopeÂ	449935665
TitanicÂ	458672302

Movie with Highest Profit Margin	Amount
AvatarÂ	523505847

MOVİE DURATİON ANALYSİS:



Key Findings from IMDB Movie Analysis

Genre Impact on IMDB Scores:

- **Most Common Genres:** Action, Drama, Comedy, and Thriller are the dominant genres in the dataset.
- **Descriptive Statistics by Genre:**
  - Action and Drama genres have relatively higher average IMDB scores compared to Comedy and Romance.
- **Insight:** Genres like Action and Drama tend to have more consistent positive reviews, while Comedy and Romance see a broader range of opinions.

Duration vs IMDB Scores:

Correlation: There is a weak positive correlation between movie duration and IMDB score, suggesting that longer movies do not necessarily lead to higher ratings.

Excel File:

https://docs.google.com/spreadsheets/d/1J6XKIpgtr5qcg0dEN2KIeFg8kCreyxTk/edit?usp=sharing&ouid=103428047773693985368&rtpof=true&sd=true

Findings

Language Analysis and IMDB Scores:

- **Most Common Languages:** English, Spanish, French, and Hindi are the most prevalent languages in the dataset.
- **IMDB Score Distribution by Language:**

English-language movies generally have the highest average IMDB score , followed by French and Spanish

Insights:

A director's track record and established reputation play a significant role in movie success, suggesting that high-profile directors attract positive reviews and larger audiences.

Findings

Director Analysis:

Top Directors:

Directors like J**ohn Blanchard, Sadyk Sher-Niyaz , Mitchell Altieri** consistently produced movies with **higher-than-average IMDB scores.**

Insights:

- **Insight: A director's track record and established reputation play a significant role in movie success, suggesting that high-profile directors attract positive reviews and larger audiences.**

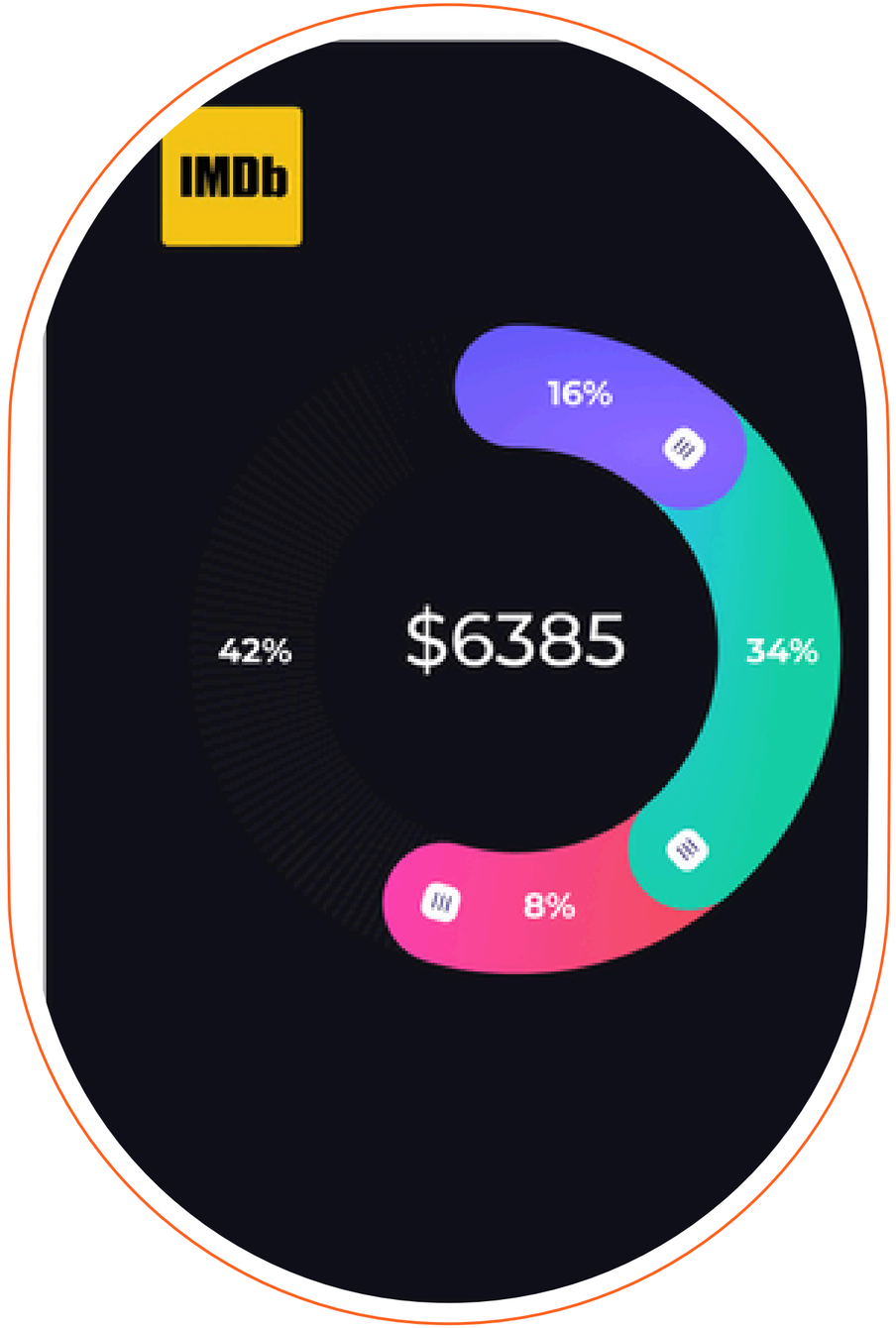
Findings

Budget and Financial Success:

Correlation: There is a moderate positive correlation between movie budget and gross earnings, suggesting that higher budgets tend to lead to higher earnings.

Insight:

While big-budget films generally perform better financially, smaller-budget films with strong storytelling or niche appeal can generate significant returns. Budget allocation should focus on both high-impact elements (e.g., cast, director) and cost-effective strategies (e.g., marketing).



Thank you

Link to Video Explanation

Connect with me on:
LinkedIn