

IMDB MOVIE ANALYSIS

By Renuka Joshi

DESCRIPTION

The IMDb Movie Analysis project aims to explore and analyze a comprehensive dataset of movies available on the IMDb platform. Through in-depth data analysis using Excel, Data Visualization and Statistics techniques this project seeks to extract valuable insights and trends that contribute to a movie's success.

The major goal of this project is to identify key factors that contribute to a movie's success measured by it's IMDb ratings. Understanding these factors can help movie directors, producers and investors make data driven decisions.

APPROACH

- **Data Cleaning:** It includes identifying & handling missing values, removing duplicates and refining the dataset.
- **Data analysis:** Evaluating various factors such as genre, movie duration, language, director impact & budget influence on ratings.
- **Findings and insights:** Summarizing trends and relationships between variables.
- **Final report:** Documenting the results with visualizations.

Microsoft Excel: Used for data cleaning, analysis, and visualization.

DATASET OVERVIEW

The key columns in the dataset include:

- **movie_title:** The name of the movie.
- **Imdb_score:** The average rating given by users on IMDB.
- **genres:** The category or categories the movie belongs to (e.g., Action, Drama, Comedy).
- **duration:** The length of the movie in minutes.
- **language:** The primary language in which the movie was released.
- **country:** The primary region where the movie was released.
- **director_name:** The name of the director responsible for the movie.
- **budget:** The estimated production budget in USD.
- **gross:** The total revenue generated by the movie worldwide.

DATA CLEANING



DATA REDUCTION

The process involved identifying and removing columns that were not relevant to the project and did not provide meaningful insights. After refinement, the dataset was reduced to nine essential columns: director's name, duration, movie title, genre, budget, gross, IMDB rating, language, and country.



HANDLING MISSING VALUES

Missing entries were addressed by locating and removing blank rows using Excel's "Go to Special" tool and the "CTRL + -" command. This refined the dataset from 5,043 to 3,888 rows, preserving only complete records for analysis.



HANDLING DUPLICATE VALUES

Column movie_title has 105 duplicate values. They should either be removed or replaced with correct value.

**THE CLEANED DATASET IS PROVIDED BELOW
FOR REFERENCE:**

DATA ANALYSIS



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.



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.

DATA ANALYSIS



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.



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.

DATA ANALYSIS



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.

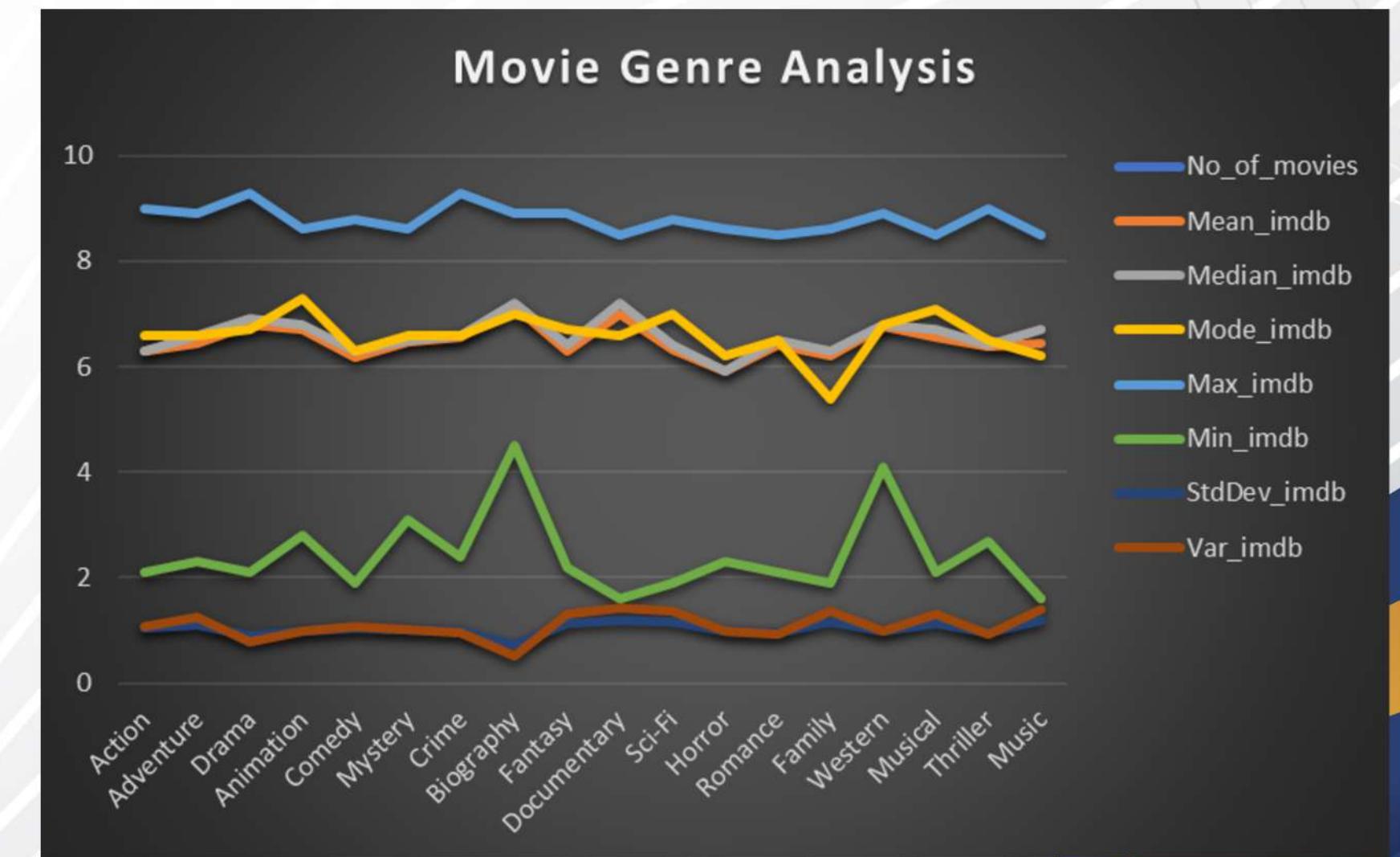
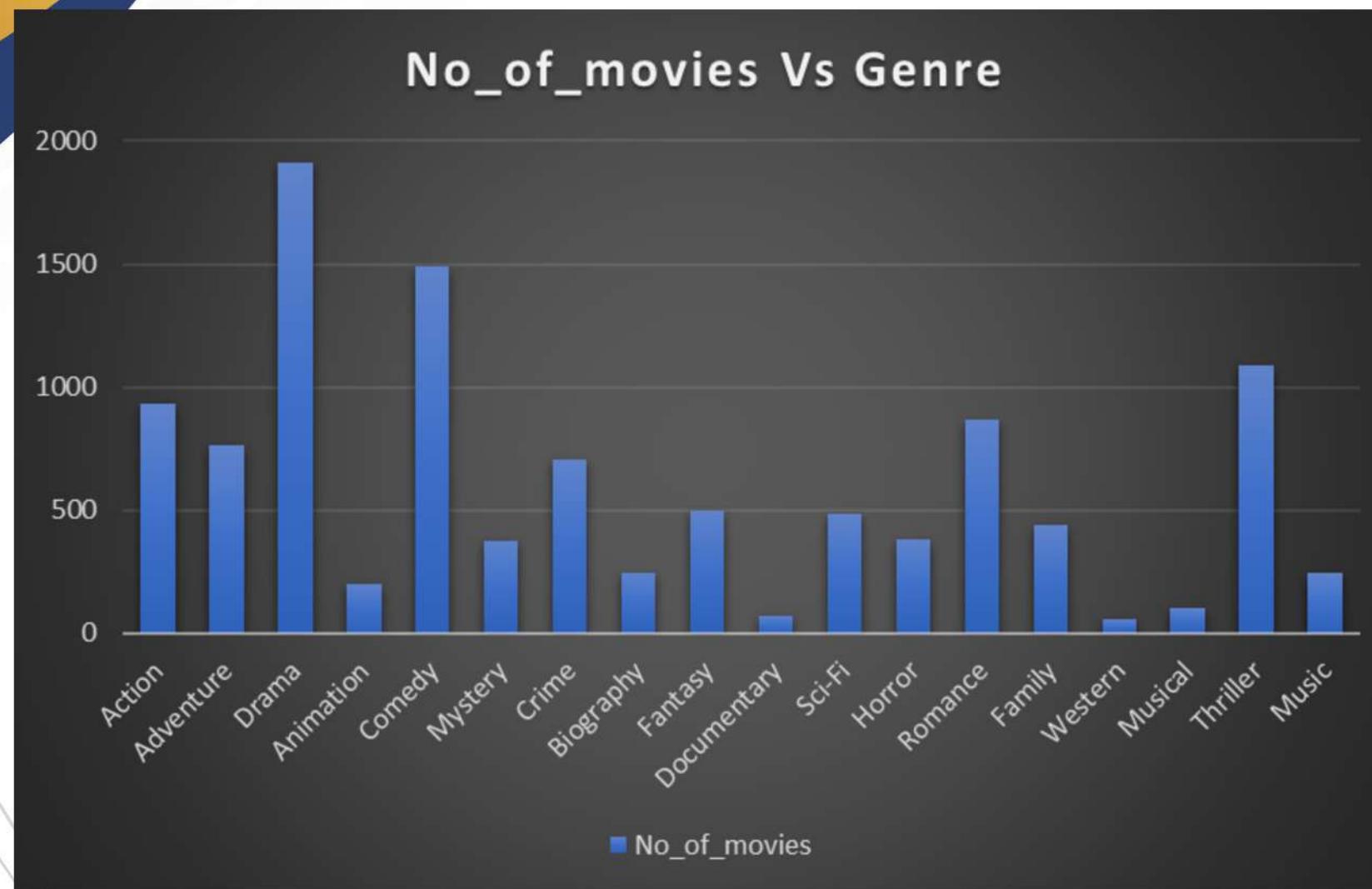
THE CLEANED DATASET LINK:-

https://github.com/RenukaJoshi-DA/IMDB_Movie_Analysis_Excel/blob/main/CleanedDataset.xlsx

MOVIE GENRE ANALYSIS

Genre	No_of_movies	Mean_imdb	Median_imdb	Mode_imdb	Max_imdb	Min_imdb	StdDev_imdb	Var_imdb
Action	935	6.2860	6.3	6.6	9	2.1	1.03780	1.0770
Adventure	766	6.4550	6.6	6.6	8.9	2.3	1.11620	1.2459
Drama	1910	6.7890	6.9	6.7	9.3	2.1	0.89105	0.7940
Animation	197	6.7005	6.8	7.3	8.6	2.8	0.99110	0.9823
Comedy	1491	6.1828	6.3	6.3	8.8	1.9	1.03970	1.0810
Mystery	377	6.4695	6.5	6.6	8.6	3.1	1.00605	1.0121
Crime	702	6.5481	6.6	6.6	9.3	2.4	0.98340	0.9671
Biography	242	7.1401	7.2	7	8.9	4.5	0.70863	0.5022
Fantasy	496	6.2851	6.4	6.7	8.9	2.2	1.13926	1.2979
Documentary	67	7.0119	7.2	6.6	8.5	1.6	1.19095	1.4184
Sci-Fi	483	6.3259	6.4	7	8.8	1.9	1.16678	1.3614
Horror	378	5.9011	5.9	6.2	8.6	2.3	0.98941	0.9789
Romance	866	6.4262	6.5	6.5	8.5	2.1	0.96844	0.9379
Family	441	6.2000	6.3	5.4	8.6	1.9	1.16825	1.3648
Western	58	6.7655	6.8	6.8	8.9	4.1	0.98987	0.9798
Musical	102	6.5510	6.7	7.1	8.5	2.1	1.13792	1.2949
Thriller	1087	6.3723	6.4	6.5	9	2.7	0.96863	0.9382
Music	247	6.4567	6.7	6.2	8.5	1.6	1.18644	1.4076

MOVIE GENRE ANALYSIS



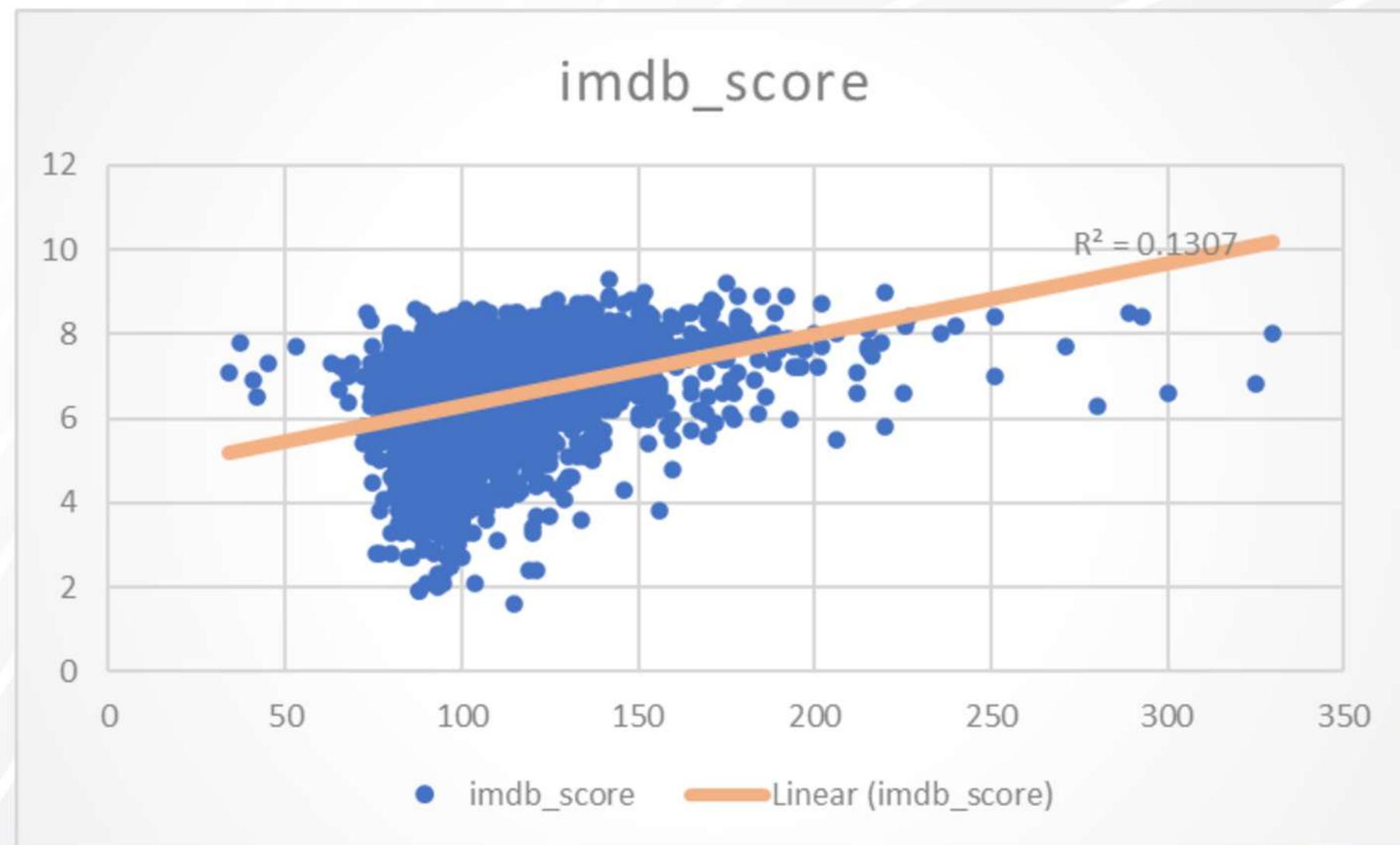
MOVIE GENRE ANALYSIS

INSIGHTS:

- Action, Drama, Comedy and Thriller are the most frequent genres.
- Drama movies had the highest average IMDB rating.
- Documentary ($\text{Var_imdb} = 1.4184$) and Music ($\text{Var_imdb} = 1.4076$) have the highest rating variance, indicating that audiences have widely different opinions on these genres.

MOVIE DURATION ANALYSIS

Operations	Values
Mean	109.81
Median	105.00
Mode	101.00
Standard Deviation	22.76
Variance	518.16



MOVIE DURATION ANALYSIS

INSIGHTS:

- ➊ The Average duration of a Movie is 109 minutes.
- ➋ The trendline between the duration vs imdb score is elevated upward with $R^2 = 0.131$
- ➌ A weak positive correlation was observed between duration and IMDB rating.

LANGUAGE ANALYSIS

Language	No_of_movies	Average_imdb	Median_imdb	Var_imdb	StdDev_imdb
English	3606	6.4214	6.5000	1.1074	1.0524
Mandarin	14	7.0214	7.2500	0.5445	0.7379
Aboriginal	2	6.9500	6.9500	0.3025	0.5500
Spanish	26	7.0500	7.1500	0.6563	0.8102
French	37	7.2865	7.2000	0.3066	0.5537
Filipino	1	6.7000	6.7000	0.0000	0.0000
Maya	1	7.8000	7.8000	0.0000	0.0000
Kazakh	1	6.0000	6.0000	0.0000	0.0000
Telugu	1	8.4000	8.4000	0.0000	0.0000
Cantonese	8	7.2375	7.3000	0.1698	0.4121
Japanese	12	7.6250	7.8000	0.7419	0.8613
Aramaic	1	7.1000	7.1000	0.0000	0.0000
Italian	7	7.1857	7.0000	1.1441	1.0696
Dutch	3	7.5667	7.8000	0.1089	0.3300
Dari	2	7.5000	7.4000	0.5028	0.7091
German	13	7.6923	7.7000	0.3792	0.6158
Mongolian	1	7.3000	7.3000	0.0000	0.0000
Thai	3	6.6333	6.6000	0.1356	0.3682
Bosnian	1	4.3000	4.3000	0.0000	0.0000
Korean	4	7.8750	7.9000	0.1719	0.4146
Hungarian	1	7.1000	7.1000	0.0000	0.0000
Hindi	10	6.7600	7.0500	1.1124	1.0547
Icelandic	1	6.9000	6.9000	0.0000	0.0000
Danish	3	7.9000	8.1000	0.1867	0.4320
Portuguese	5	7.7600	8.0000	0.7664	0.8754
Norwegian	4	7.1500	7.3000	0.2475	0.4975
Czech	1	7.4000	7.4000	0.0000	0.0000
Russian	1	6.5000	6.5000	0.0000	0.0000
None	1	8.5000	8.5000	0.0000	0.0000
Zulu	1	7.3000	7.3000	0.0000	0.0000
Hebrew	3	7.5000	7.3000	0.1267	0.3559
Dzongkha	1	7.5000	7.5000	0.0000	0.0000
Arabic	1	7.2000	7.2000	0.0000	0.0000
Vietnamese	1	7.4000	7.4000	0.0000	0.0000
Indonesian	2	7.9000	7.9000	0.0900	0.3000
Romanian	1	7.9000	7.9000	0.0000	0.0000
Persian	3	8.1333	8.4000	0.2022	0.4497
Swedish	1	7.6000	7.6000	0.0000	0.0000

LANGUAGE ANALYSIS

INSIGHTS:

- The Most common languages used in the movies are English, French, Spanish, Mandarin and German.
- English-language movies have the largest volume but not the highest ratings.
- It is also Observed that the languages Telugu and Persian have the highest average imdb score.
- Languages with fewer movies tend to have either extreme high or low ratings due to limited data.

DIRECTOR ANALYSIS

director_name	No_of_Movies	Average_IMDB_Score	Top_Director_Classification	Rank	90th_Percentile
James Cameron	7	7.91	Top Director	41	7.5
Gore Verbinski	7	6.99	Below Percentile	487	
Sam Mendes	7	7.46	Below Percentile	190	
Christopher Nolan	8	8.43	Top Director	8	
Andrew Stanton	3	7.73	Top Director	80	
Sam Raimi	10	6.96	Below Percentile	494	
Nathan Greno	1	7.80	Top Director	58	
Joss Whedon	3	7.87	Top Director	54	
David Yates	3	7.20	Below Percentile	279	
Zack Snyder	7	7.14	Below Percentile	342	
Bryan Singer	8	7.29	Below Percentile	264	
Marc Forster	7	7.23	Below Percentile	277	
Andrew Adamson	4	7.15	Below Percentile	335	
Rob Marshall	5	6.60	Below Percentile	735	
Barry Sonnenfeld	7	6.46	Below Percentile	876	
Peter Jackson	9	7.89	Top Director	53	
Marc Webb	3	7.13	Below Percentile	345	
Ridley Scott	16	7.13	Below Percentile	349	
Chris Weitz	5	6.08	Below Percentile	1141	
Anthony Russo	4	7.00	Below Percentile	419	
Peter Berg	6	6.67	Below Percentile	708	
Colin Trevorrow	2	7.00	Below Percentile	419	
Shane Black	2	7.40	Below Percentile	201	
Tim Burton	14	7.05	Below Percentile	408	
Brett Ratner	9	6.46	Below Percentile	877	
Dan Scanlon	1	7.30	Below Percentile	231	
Michael Bay	12	6.62	Below Percentile	730	

Top Director's List With Their Rank

director_name	No_of_Movies	Average_IMDB_Score	Rank
James Cameron	7	7.91	41
Christopher Nolan	8	8.43	8
Andrew Stanton	3	7.73	80
Nathan Greno	1	7.80	58
Joss Whedon	3	7.87	54
Peter Jackson	9	7.89	53
Lee Unkrich	1	8.30	13
Steven Spielberg	25	7.54	159
Pete Docter	3	8.23	17
Martin Scorsese	16	7.68	107
Brad Bird	5	7.58	151
Don Hall	1	7.90	42
Rich Moore	1	7.80	58
Dean DeBlois	3	7.77	74
David Fincher	10	7.75	79
Matthew Vaughn	4	7.65	110
Duncan Jones	3	7.57	153
Chris Buck	1	7.60	122
Peter Weir	4	7.73	82
Alejandro G. Iñárritu	5	7.84	56
Paul Greengrass	7	7.59	150
Mark Osborne	2	7.70	105
Alfonso Cuarón	4	7.80	58
Quentin Tarantino	8	8.20	19
Spike Jonze	4	7.58	152

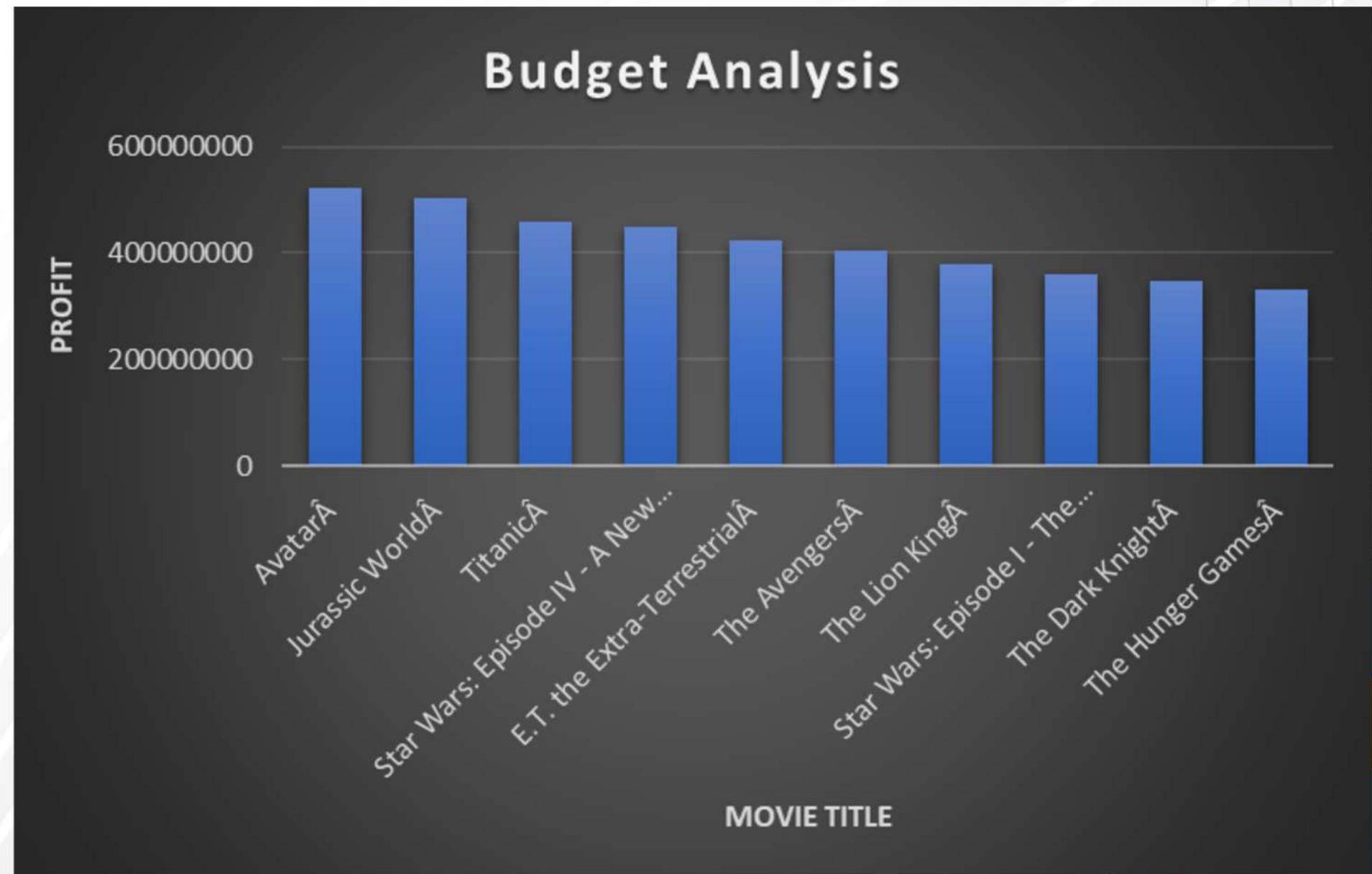
DIRECTOR ANALYSIS

INSIGHTS:

- It is observed that Tony Kaye, Charles Chaplin, Alfred Hitchcock, Ron Fricke, Damien Chazelle, Majid Majidi, Sergio Leone, Christopher Nolan, SS Rajamouli and Richard Marquand are the top 10 directors with average imdb score ≥ 8.4
- Top directors (90th percentile = 7.5) had consistently high ratings across multiple films.
- Lesser-known directors had more varied ratings, possibly indicating inconsistency in quality.

BUDGET ANALYSIS

movie_title	Profit
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A New HopeÂ	449935665
E.T. the Extra-TerrestrialÂ	424449459
The AvengersÂ	403279547
The Lion KingÂ	377783777
Star Wars: Episode I - The Phantom MenaceÂ	359544677
The Dark KnightÂ	348316061
The Hunger GamesÂ	329999255



Correlation between budget and Gross earnings

0.223115594

BUDGET ANALYSIS

INSIGHTS:

- The Top-5 with highest profits are Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope and E.T. The Extra Terrestrial.
- The Correlation between budget and gross is positive.
- High-budget movies generally had better financial success but not necessarily higher IMDB ratings.

THE RESULT DATASET LINK :

[https://github.com/RenukaJoshi-
DA/IMDB_Movie_Analysis_Excel/blob/main/IMDB_Movie_Analysis.xlsx](https://github.com/RenukaJoshi-DA/IMDB_Movie_Analysis_Excel/blob/main/IMDB_Movie_Analysis.xlsx)

CONCLUSION

The analysis provides actionable insights for movie producers, directors, and investors to make data-driven decisions for film production and marketing strategies.

- Genres like Drama and Action tend to receive higher ratings, while Horror movies show inconsistent audience reception.
- Ideal movie duration lies between 90-120 minutes; excessively long or short films tend to receive lower ratings.
- The Most common languages used in the movies are English, French, Spanish, Mandarin and German. But non-English films exhibit greater variability.
- Directors in the 90th percentile consistently produce high-rated movies, highlighting the importance of experienced filmmakers.
- Budget strongly correlates with earnings, but does not guarantee a higher IMDB rating.

THANK YOU