IMDB Movies Analysis

Project Description

This project aims to analyze a dataset of IMDB movies to uncover meaningful insights regarding movie genres, durations, languages, directors, and budgetary impact on financial success. The primary objectives include identifying trends in movie ratings, understanding the influence of language and director choices, and analyzing budget-to-revenue relationships to determine profitability. The approach involves statistical analysis and visualization techniques to extract key patterns from the data.

Approach

The project was executed using a structured data analysis pipeline:

- ➢ **Data Cleaning & Preprocessing**
  - Loading the dataset and removing unnecessary columns.
  - Handling missing values and standardized categorical data.
  - Saving the cleaned dataset with the name of NEW_IMDB_Movies.xlsx using python (Jupyter Notebook).
- ➢ **Movie Genre Analysis**
  - Identified the most common genres and their impact on IMDB scores.
  - Used descriptive statistics (mean, median, mode, range, variance, standard deviation) for analysis.
- ➢ **Movie Duration Analysis**
  - Examined the distribution of movie durations.
  - Assessed relation between duration and IMDB scores using scatter plots and trend analysis.
- ➢ **Language Analysis**
  - Determined the most common movie languages.
  - Calculated descriptive statistics to evaluate language influence on IMDB ratings.
- ➢ **Director Analysis:**
  - Identified top directors based on their average IMDB scores.
  - Used percentile calculations to rank top-performing directors.
- ➢ **Budget Analysis:**
  - Analyzed correlation between budget and financial success.
  - Identified movies with the highest profit margins.

Tech-Stack Used

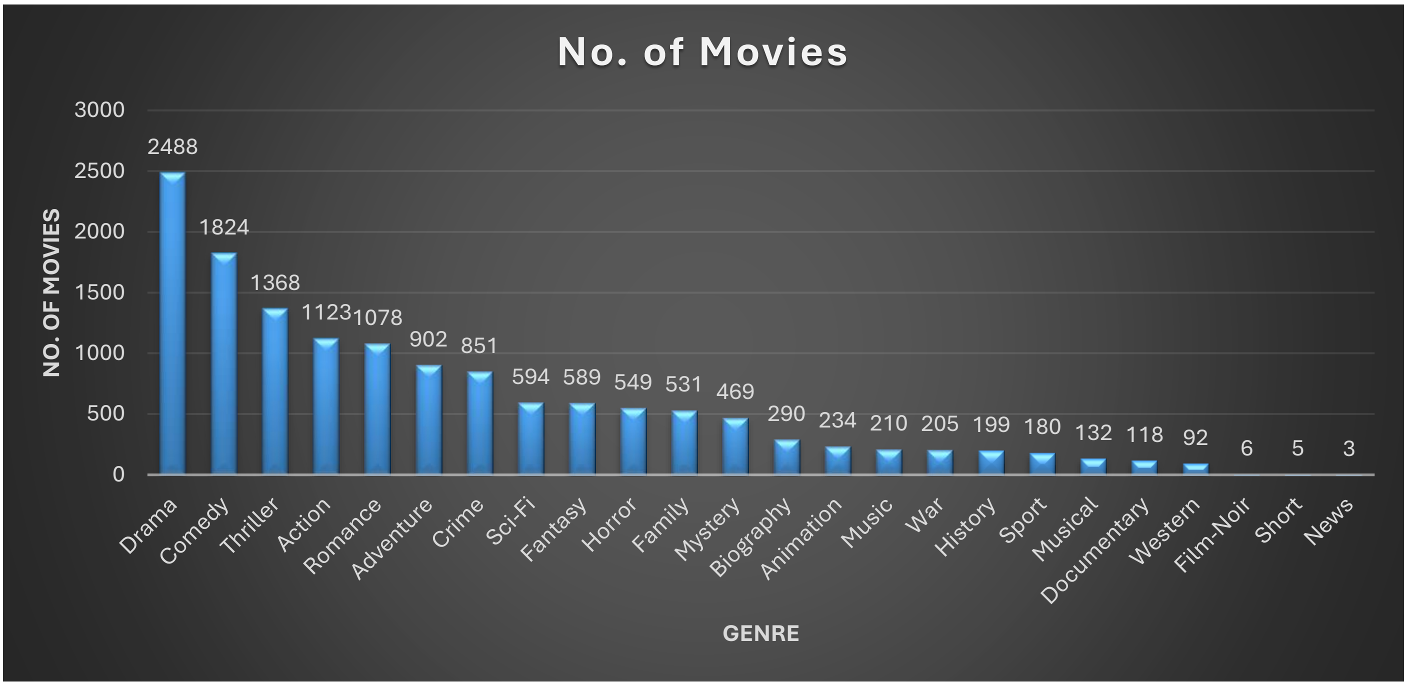
In this project, we have used-

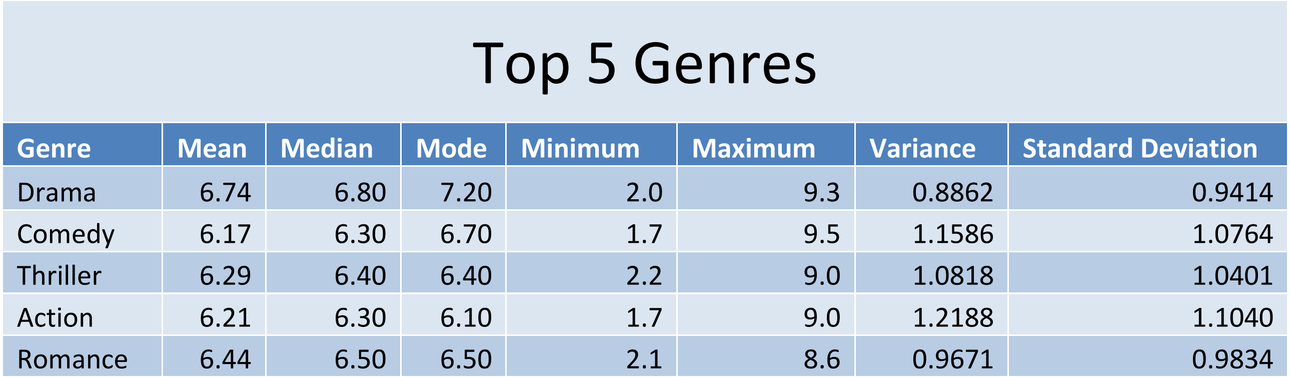
- Python (Jupyter Notebook, Pandas, NumPy) – Employed for advanced data cleaning, analysis, and visualization of trends.
- Microsoft Excel 2024 – Used for initial data exploration, applying formulas (COUNTIF, AVERAGE, MEDIAN, STDEV, CORREL) for statistical calculations, using graphs for visualization and summarizing data.
- Google Drive – Hosting and sharing reports.

Data Analytics Tasks:

Movie Genre Analysis:

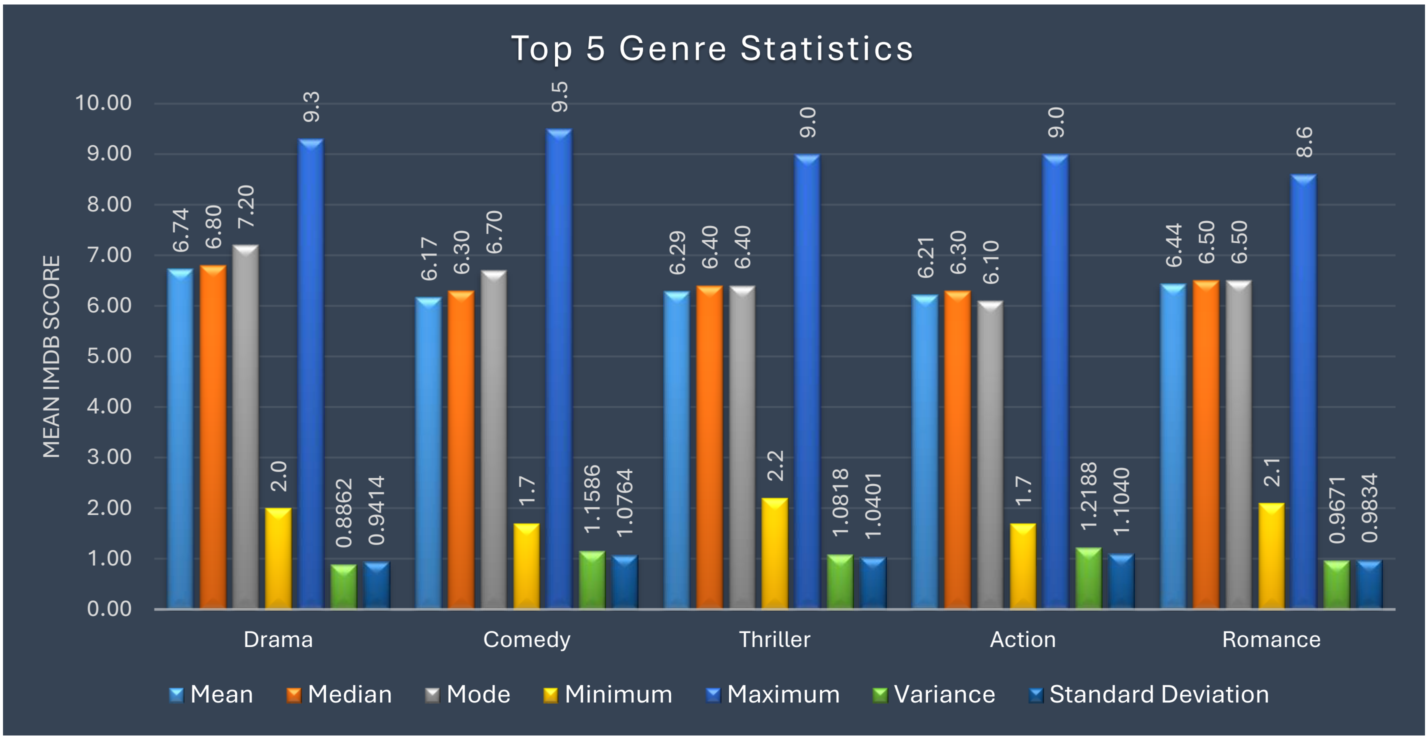
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation) of the IMDB scores.





Top 5 Genres

Genre	Mean	Median	Mode	Minimum	Maximum	Variance	Standard Deviation
Drama	6.74	6.80	7.20	2.0	9.3	0.8862	0.9414
Comedy	6.17	6.30	6.70	1.7	9.5	1.1586	1.0764
Thriller	6.29	6.40	6.40	2.2	9.0	1.0818	1.0401
Action	6.21	6.30	6.10	1.7	9.0	1.2188	1.1040
Romance	6.44	6.50	6.50	2.1	8.6	0.9671	0.9834

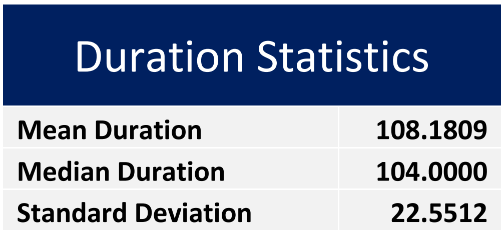


Observations:

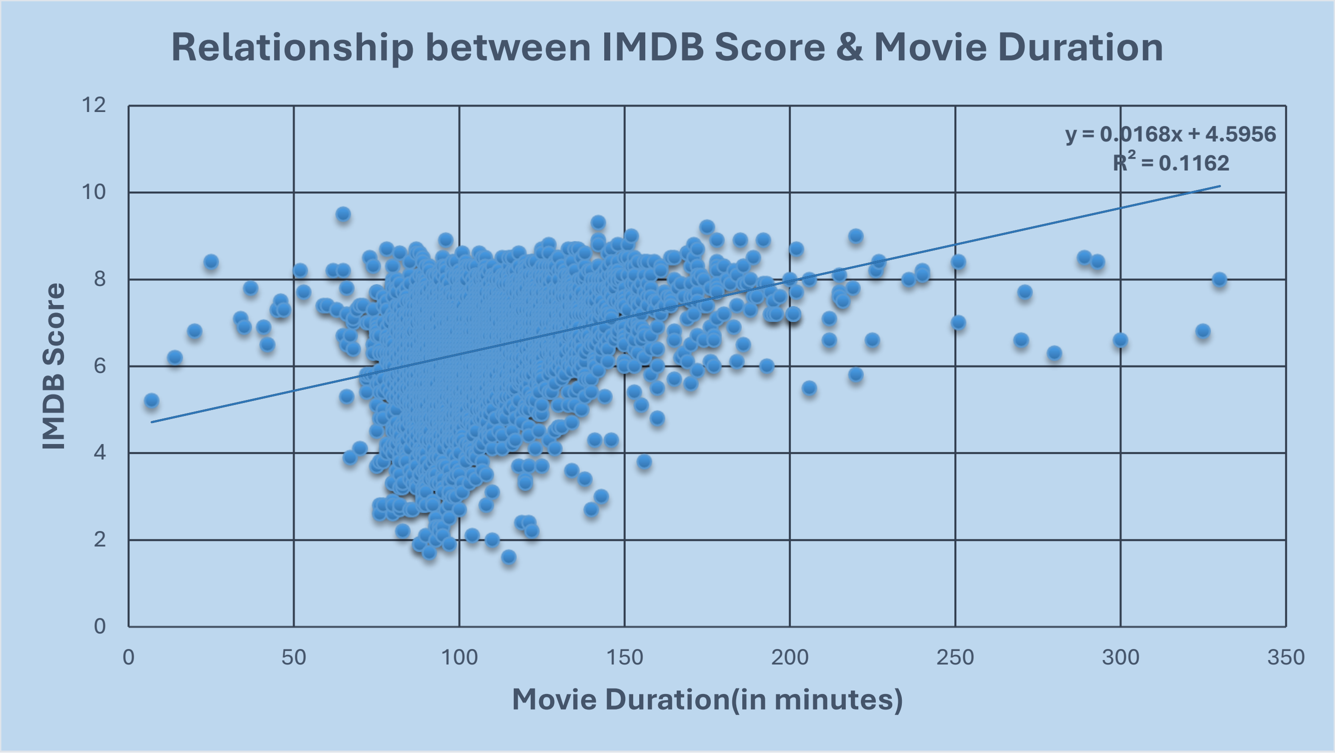
- Despite being very popular Genres like Drama, Comedy, Thriller etc. their mean IMDB score is low but, Movies with highest IMDB rating range from 9.0-9.5 are belongs from these genres which means they are mass-produced and vary in quality, impacting the mean.
- Genres with lowest standard deviation like Film-Noir and News with highest average IMDB ratings are also most consistent in ratings which means very little fluctuation in quality.
- If a producer wants to choose a genre with high ratings and good audience reception, then, they can choose genre like Biography, history, war, Documentary because of their good average rating and lower deviation.

Movie Duration Analysis:

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



Duration Statistics	
Mean Duration	**108.1809**
Median Duration	**104.0000**
Standard Deviation	**22.5512**



Observations:

- The trendline suggests a slight positive correlation, meaning longer movies tend to have slightly higher IMDB ratings.
- However, the variation is high, meaning other factors also influence IMDB ratings significantly.

Language Analysis: Situation:

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

Top 5 Languages

Language	No. of Movies	mean	median	Standard Deviation
English	4567	6.37	6.50	1.1111
French	72	7.02	7.20	0.7114
Spanish	40	6.94	7.15	0.8443
Hindi	28	6.63	6.95	1.3737
Mandarin	24	6.79	7.05	1.0150

Top 5 Language as per mean IMDB Score

Language	No. of Movies	mean	median	Standard Deviation
Telugu	1	8.40	8.40	0.0000
Indonesian	2	7.90	7.90	0.3000
Maya	1	7.80	7.80	0.0000
Hebrew	5	7.58	7.60	0.2993
Persian	4	7.58	7.95	1.0425

Observations:

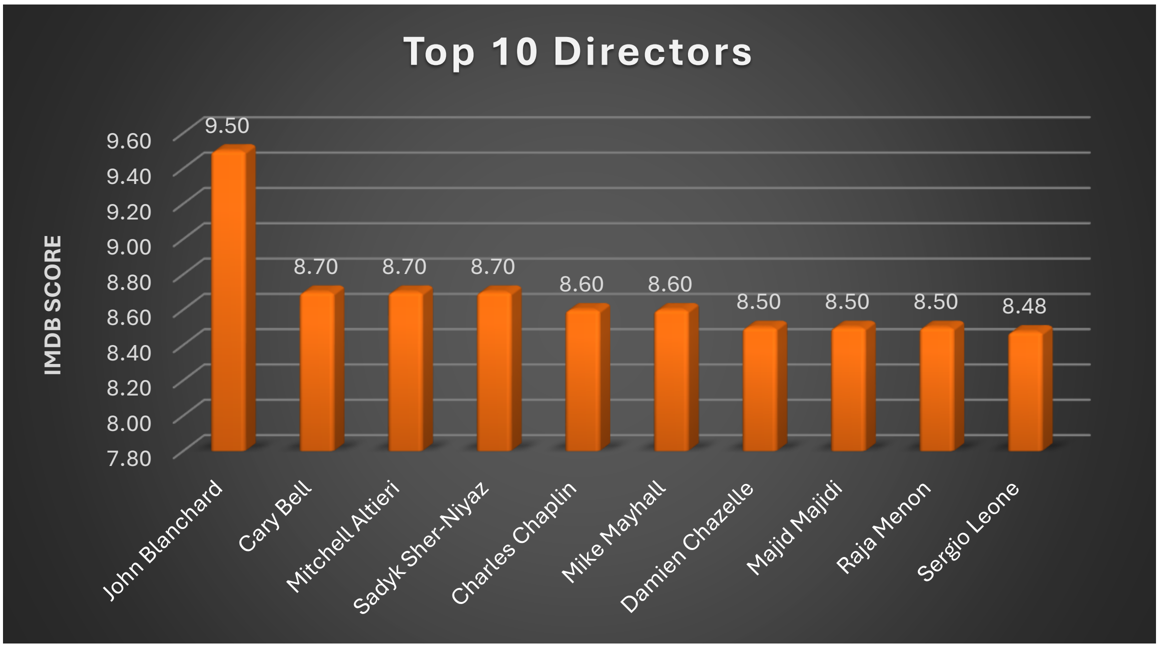
- **English** is the most common language in movies which dominate in quantity. Also, **Hindi** have good numbers of movies but their average ratings are **low** which is possibly **due to high deviation** in IMDB ratings of movies.
- **French, Spanish** tends to have **better average IMDB ratings** than English movies with **lower deviation**.
- **Telugu** has the highest average rating i.e. **8.4** (but only 1 movie).
- Language with fewer movies tends to have higher average IMDB ratings due to their limited data points while language like English, French, Spanish with large number of movies tends to have lower average IMDB ratings due to high standard deviation in ratings.

Director Analysis:

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

Using percentile calculations, we identified directors whose average IMDB scores place them in the top 5% (above the 95th percentile). The highest-rated directors include:

Top 10 Directors whose score is more than 95th percentile i.e. 7.8	
John Blanchard	9.50
Cary Bell	8.70
Mitchell Altieri	8.70
Sadyk Sher-Niyaz	8.70
Charles Chaplin	8.60
Mike Mayhall	8.60
Damien Chazelle	8.50
Majid Majidi	8.50
Raja Menon	8.50
Sergio Leone	8.48



Observations:

- These directors consistently produce highly-rated films, with **John Blanchard** leading at **9.5.**
- Well-known directors like **Charlie Chaplin, Sergio Leone** also feature among the best.
- The presence of contemporary directors (e.g., **Damien Chazelle**) suggests that top-rated films span different eras.

Budget Analysis:

Task: Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

Correlation between Movie Budget & Gross earning is	0.1001842

Top 10 Movies with Highest Profit Margin	
Avatar	$ 523,505,847.00
Jurassic World	$ 502,177,271.00
Titanic	$ 458,672,302.00
Star Wars: Episode IV - A New Hope	$ 449,935,665.00
E.T. the Extra-Terrestrial	$ 424,449,459.00
The Avengers	$ 403,279,547.00
The Lion King	$ 377,783,777.00
Star Wars: Episode I - The Phantom Menace	$ 359,544,677.00
The Dark Knight	$ 348,316,061.00
The Hunger Games	$ 329,999,255.00

Observations:

- Since 0.1001842 is close to 0, the relationship between budget and gross earnings is very weak. This means higher budgets do not strongly predict higher earnings.
- Because r is positive, it suggests that as budget increases, gross earnings tend to increase slightly. However, this effect is very small.

Insights

- Popular genres (Drama, Comedy, Thriller) have low average ratings due to mass production and inconsistent quality.
- Biography, History, War, Documentary genres have high average ratings and low variation, making them ideal for quality content.
- Film-Noir & News genres are most consistent with high ratings, though niche.
- Longer movies slightly score higher, but other factors impact ratings more.
- English dominates, but French & Spanish films have better average ratings.
- Top-rated directors like *John Blanchard* and *Charlie Chaplin* deliver consistently high-quality films.
- Budget and earnings are weakly correlated (r ≈ 0.10) → High budget ≠ high earnings.

Results

The project successfully identified patterns and insights within the IMDb dataset. It enhanced understanding of:

- What genres and languages consistently perform well
- How duration and director influence ratings
- Why financial success doesn't always align with production cost

Useful links

To know how we clean dataset using python (Jupyter notebook) click below link (kindly download file to view)-

Dataset cleaning using python

To view cleaned dataset and analysis file (recommended to view in Microsoft excel)-

Cleaned dataset