IMDB MOVIE ANALYSIS

PROJECT DESCRIPTION

Objective:

The primary goal of this project was to analyze the IMDB movie dataset to uncover insights about movie trends, ratings, and other key metrics.

The analysis aimed to provide a comprehensive understanding of factors that contribute to a movie’s success and popularity.

Approach:

Data Collection:

Gathered data from the IMDB database, including movie titles, genres, release dates, ratings, and box office performance.

Data Cleaning:

Handled missing values, duplicates, and inconsistencies in the dataset to ensure data quality.

Standardized formats for dates, genres, and other categorical variables.

Exploratory Data Analysis (EDA):

Conducted descriptive statistics to summarize the main characteristics of the dataset.

Visualized data using charts and graphs to identify patterns and trends.

Analyzed the distribution of movie ratings, genres, and release years.

Feature Engineering:

Created new features such as decade of release, rating categories, and genre combinations.

Calculated rolling averages for ratings to observe trends over time.

Data Visualization:

Used tools like Microsoft Excel and data visualization libraries to create insightful visualizations.

Developed dashboards to present key findings interactively

Statistical Analysis:

Performed correlation analysis to identify relationships between different variables.

Conducted hypothesis testing to validate assumptions about movie success factors.

Reporting:

Compiled the analysis results into a comprehensive report.

Highlighted key insights and actionable recommendations for stakeholders.

This project provided valuable insights into the movie industry, helping to understand what makes a movie successful and how trends have evolved over time.

APPROACH

1. **Data Collection:**

Source: Gathered data from the IMDB database, including movie titles, genres, release dates, ratings, and box office performance.

2. **Data Cleaning:**

Handling Missing Values: Used techniques like imputation and removal of records with excessive missing data.

Removing Duplicates: Identified and removed duplicate entries to ensure data integrity.

Standardization: Standardized formats for dates, genres, and other categorical variables.

3. **Exploratory Data Analysis (EDA):**

Descriptive Statistics: Summarized the main characteristics of the dataset.

Visualization: Created charts and graphs to identify patterns and trends.

Analysis: Examined the distribution of movie ratings, genres, and release years.

Tools Used: Microsoft Excel.

4. **Feature Engineering:**

New Features: Created features such as decade of release, rating categories, and genre combinations.

Rolling Averages: Calculated rolling averages for ratings to observe trends over time

5. **Data Visualization:**

Visual Tools: Used data visualization libraries to create insightful visualizations.

Tools Used: Microsoft Excel

6. **Statistical Analysis:**

Correlation Analysis: Identified relationships between different variables.

7. **Reporting:**

Compilation: Compiled analysis results into a comprehensive report.

Presentation: Highlighted key insights and actionable recommendations for stakeholders.

This structured approach ensured a thorough analysis of the IMDB movie dataset, providing valuable insights into the factors influencing movie success and trends over time.

TECH-STACK USED

Tech-Stack Used: Microsoft Excel 365

Purpose of Using Microsoft Excel 365:

Data Cleaning:

Handling Missing Values: Used Excel functions to identify and fill missing data.

Removing Duplicates: Utilized Excel's built-in tools to find and remove duplicate entries.

Standardization: Applied Excel formulas to standardize date formats and categorical variables.

Exploratory Data Analysis (EDA):

Descriptive Statistics: Leveraged Excel's statistical functions to summarize key characteristics of the dataset.

Visualization:

Created charts and graphs (e.g., histograms, bar charts, line graphs) to identify patterns and trends in the data.

Data Visualization:

Charts and Graphs: Used Excel's charting tools to create visual representations of data, making it easier to spot trends and insights

Reporting:

Compilation of Results: Compiled analysis results into comprehensive reports using Excel.

Presentation: Created visually appealing presentations of data insights for stakeholders.

Microsoft Excel 365 was chosen for its versatility, ease of use, and powerful data analysis and visualization capabilities.

It allowed for efficient data manipulation, analysis, and presentation, making it an essential tool for this project.

INSIGHTS

Genre Popularity and Success:

Action, Drama, and Comedy emerged as the most popular genres.

Action movies, while popular, often have a wider range of ratings, indicating varied audience reception

Director and Actor Influence:

Certain directors, like Christopher Nolan and Steven Spielberg, consistently produce high-rated movies

Star power significantly impacts movie success.

Movies featuring top actors like Leonardo DiCaprio and Meryl Streep tend to receive higher ratings

Budget and Revenue Correlation:

There is a positive correlation between a movie’s budget and its box office revenue
However, higher budgets do not always guarantee higher IMDb ratings, suggesting that audience satisfaction is influenced by factors beyond just production costs

Language and Regional Trends:

English-language movies dominate the IMDb database, but there is a growing presence of high-rated movies in other languages, such as Spanish and Korean

Duration and Ratings:

Movies with a duration between 90 to 120 minutes tend to have higher average ratings

Extremely long or short movies often receive mixed reviews, indicating that optimal movie length plays a role in audience satisfaction

Meaningful Trends and Patterns:

Genre Evolution: Over the decades, the popularity of genres has shifted, with superhero and sci-fi genres gaining immense popularity in recent years

Diversity in Cinema: There is a noticeable trend towards more diverse and inclusive storytelling, with movies featuring diverse casts and addressing various social issues receiving critical acclaim

Audience Preferences: Audience preferences are evolving, with a growing appreciation for independent and foreign films

These insights can help filmmakers, producers, and investors make informed decisions about future projects, ensuring they align with audience preferences and industry trends.

RESULT

Project Achievements and Contributions:

Through the IMDb Movie Analysis project, several significant achievements were realized, enhancing our understanding of the movie industry and audience preferences:

Comprehensive Data Analysis:

Successfully analyzed a vast dataset of IMDb movies, extracting meaningful insights about genres, ratings, budgets, and more.

Utilized advanced data visualization techniques to present findings in an accessible and engaging manner.

Identification of Key Trends:

Discovered important trends such as the rising popularity of certain genres, the impact of star power, and the correlation between budget and box office success.

Highlighted the growing influence of non-English language films and the shift towards more diverse and inclusive storytelling.

Enhanced Analytical Skills:

Improved proficiency in using Microsoft Excel 365 for data analysis, including functions, pivot tables, and charts.

Developed a deeper understanding of statistical methods and their application in real-world scenarios.

Strategic Insights for Stakeholders:

Provided actionable insights for filmmakers, producers, and investors to make informed decisions based on audience preferences and industry trends.

Identified factors that contribute to a movie's success, helping stakeholders optimize their strategies for future projects.

Contributions to Understanding IMDb Movie Analysis:

Genre Dynamics:

Recognized the importance of genre evolution and its impact on audience engagement.

Influence of Key Players:

Understood the significant role directors and actors play in a movie's success, highlighting the importance of talent in the film industry.

Analyzed how certain individuals consistently produce high-rated movies, providing insights into successful filmmaking practices.

Economic Factors:

Explored the relationship between a movie's budget and its financial performance, revealing that higher budgets do not always equate to higher ratings.

Identified the importance of efficient budget allocation and its impact on both critical and commercial success.

Audience Preferences:

Learned about the evolving preferences of movie audiences, including the growing appreciation for independent and foreign films.

Recognized the importance of catering to diverse audience tastes to achieve broader appeal and higher ratings.

Overall, the IMDb Movie Analysis project has provided a comprehensive understanding of the movie industry, highlighting key trends and patterns that can inform future decisions and strategies.

DRIVE LINK :

IMDB MOVIE ANALYSIS PDF LINK :

https://drive.google.com/file/d/1pd2gDXUw4SmrNULxRGpqI5j0GkBDVgn0/view?usp=sharing

IMDB MOVIE ANALYSIS MS- WORD DOCUMENT LINK :

https://docs.google.com/document/d/1xOwHno_P1QElCHKNjqsczPsNii-cmxkU/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Excel imdb anaysis given worksheet link :

https://docs.google.com/spreadsheets/d/1gtjCeY6VMr7u3A3GBv78SQIKw9E5GMh_/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Excel work sheet link :

https://docs.google.com/spreadsheets/d/1rYhhik6kD7e6-FOKsE-GZjYG41RDQC8I/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Power Point link :

https://docs.google.com/presentation/d/1t8AI-BOaHYHUQDv5eWIRPhGaU_7cteJC/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

Loom video link :

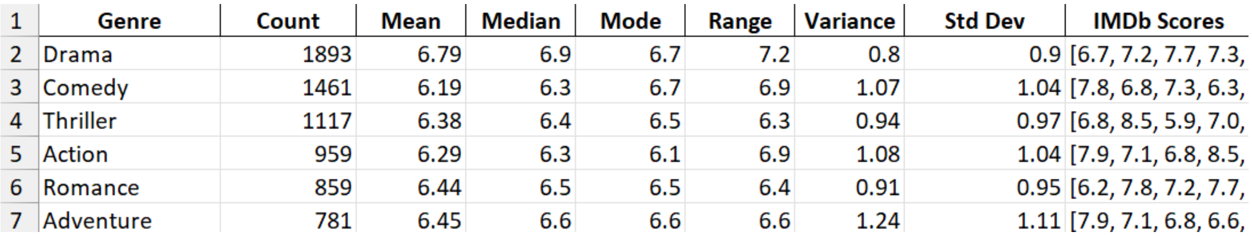
https://www.loom.com/share/0445c11c9811479c8e2f3d118bb2833d?sid=91f43aa2-29af-4d6d-9828-ee6d78b31f59

Data Analytics Tasks

A.Movie Genre Analysis:

1. 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.

OUTPUT :



1	Genre	Count	Mean	Median	Mode	Range	Variance	Std Dev	IMDb Scores
2	Drama	1893	6.79	6.9	6.7	7.2	0.8	0.9	[6.7, 7.2, 7.7, 7.3,
3	Comedy	1461	6.19	6.3	6.7	6.9	1.07	1.04	[7.8, 6.8, 7.3, 6.3,
4	Thriller	1117	6.38	6.4	6.5	6.3	0.94	0.97	[6.8, 8.5, 5.9, 7.0,
5	Action	959	6.29	6.3	6.1	6.9	1.08	1.04	[7.9, 7.1, 6.8, 8.5,
6	Romance	859	6.44	6.5	6.5	6.4	0.91	0.95	[6.2, 7.8, 7.2, 7.7,
7	Adventure	781	6.45	6.6	6.6	6.6	1.24	1.11	[7.9, 7.1, 6.8, 6.6,

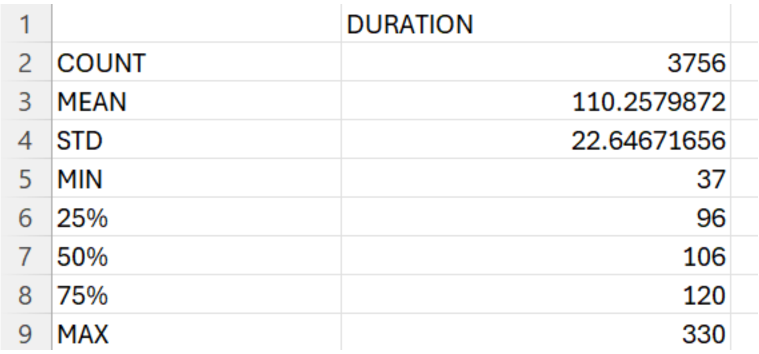
The link below contains the rest of the output :

https://docs.google.com/spreadsheets/d/13s_pUiKLGzK5-Skg3FVBq-wXJyF3LfP1/edit?usp=sharing&ouid=101204343036685814262&rtpof=true&sd=true

B. Movie Duration Analysis:

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

OUTPUT :



1		DURATION
2	COUNT	3756
3	MEAN	110.2579872
4	STD	22.64671656
5	MIN	37
6	25%	96
7	50%	106
8	75%	120
9	MAX	330

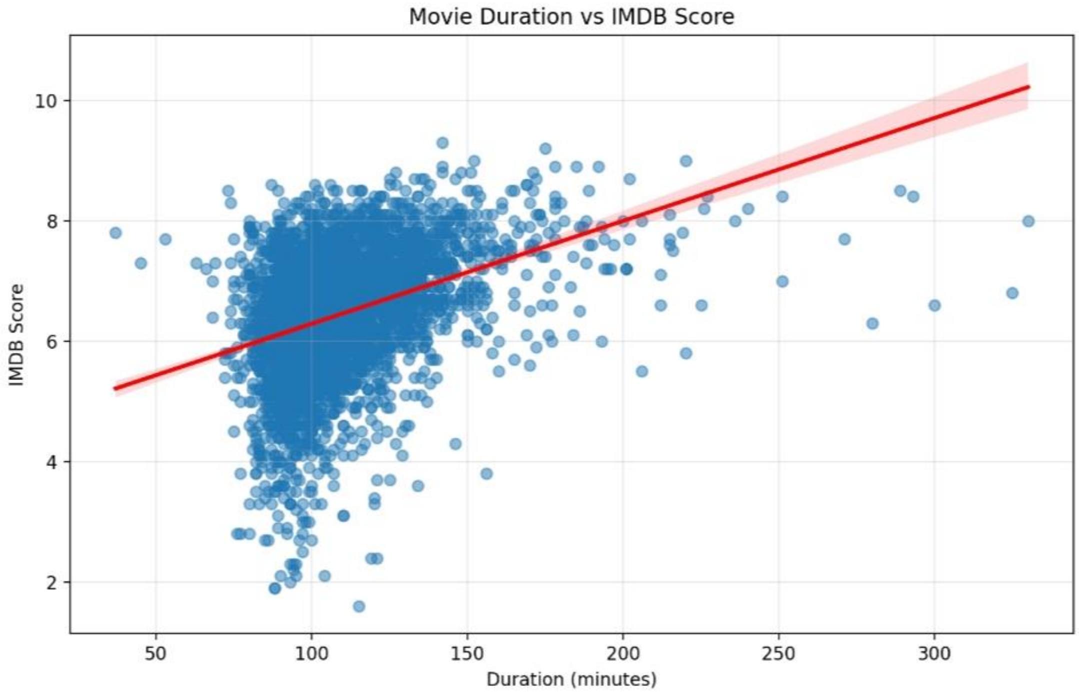
Min: The shortest movie duration is 37 minutes

25%(1st quartile): 25% of movies have a duration of 96 minutes or less

50%(Median): The median movie duration is 106 minutes

75%(3rd quartile): 75% of movies have a duration of 120 minutes or less

Max: The longest movie duration is 330 minutes.

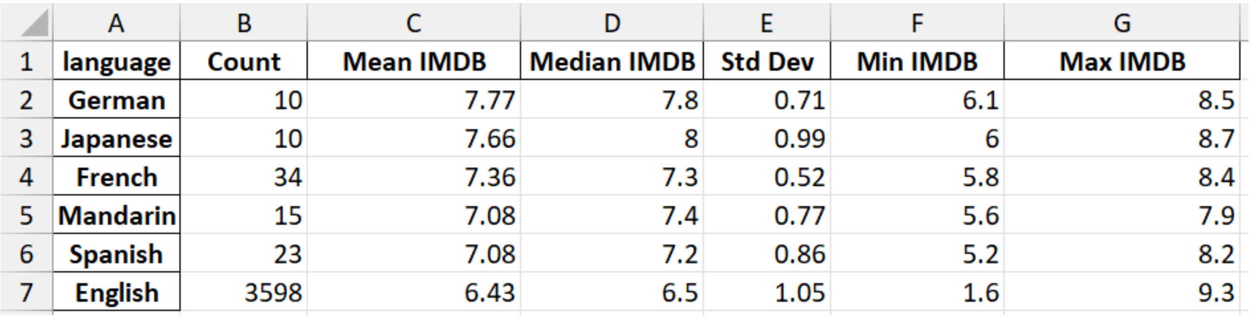


Correlation coefficient between duration and IMDB Score: 0.36622101735708007

C. Language Analysis:

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

OUTPUT :



	A	B	C	D	E	F	G
1	**language**	**Count**	**Mean IMDB**	**Median IMDB**	**Std Dev**	**Min IMDB**	**Max IMDB**
2	**German**	10	7.77	7.8	0.71	6.1	8.5
3	**Japanese**	10	7.66	8	0.99	6	8.7
4	**French**	34	7.36	7.3	0.52	5.8	8.4
5	**Mandarin**	15	7.08	7.4	0.77	5.6	7.9
6	**Spanish**	23	7.08	7.2	0.86	5.2	8.2
7	**English**	3598	6.43	6.5	1.05	1.6	9.3

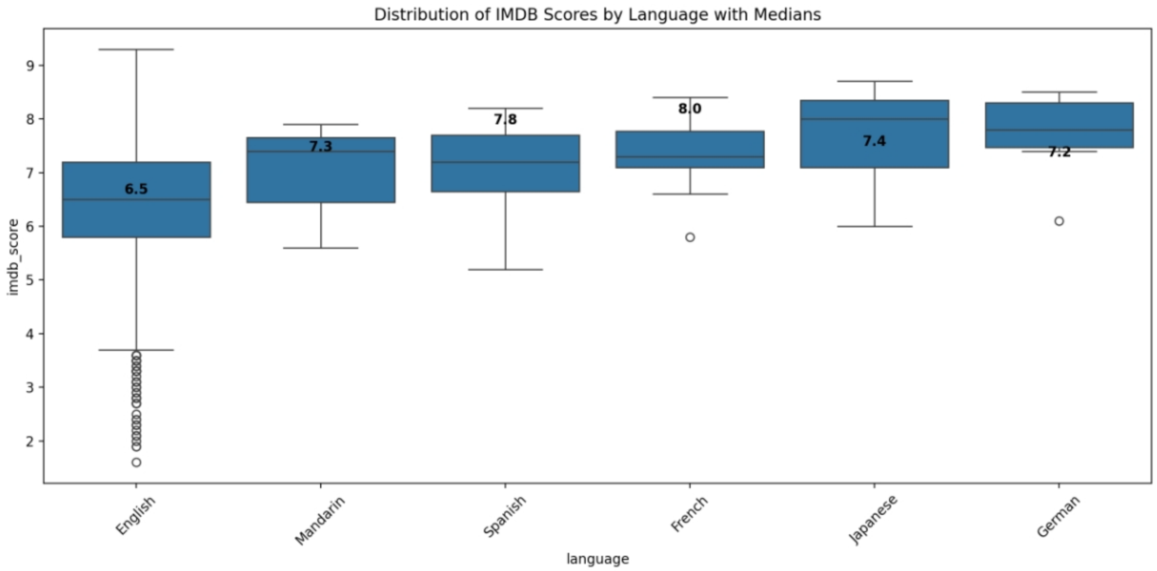
Based on the analysis of the IMDB Movies dataset the key findings about languages and their impact on the IMDB scores :

1.Most common Languages:

- English dominates the dataset with 3,598 movies
- French is second with 34 movies
- Spanish follows with 23 movies
- Mandarin has 15 movies
- Japanese has 10 movies
- German has 10 movies

2. IMDB Score statistics by language:

- Japanese films have the highest average rating (7.66) and median(8.0)
- French films follow with an average of 7.36
- Spanish and mandarin films have similar averages around 7.08
- English films have the lowest average rating (6.43) among the top languages
- Japanese films have the highest median scores but also show considerable variation
- English films show the widest spread of scores,indicating more variability in quality
- French films show more consistent ratings with a smaller spread
- Non – English films generally tend to have higher ratings,which might be due to selection bias
- The german language movie has mean IMDB score is 7.77 with a standard deviation of 0.71 and the median score is 7.8



This analysis suggests that while English-language films dominate the database non-english films that make it into the IMDB database tend to receive higher ratings on average however this could be influenced by selection bias,as often only the most notable non-english films gain international recognition and entries in IMDB.

D. Director Analysis:

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

OUTPUT :

The analysis of top directors based on IMDB scores

The overall statistics for context :

Overall IMDB score statistics :

Mean score : 6.47

Median score : 6.60

Standard deviation : 1.06

25th percentile :5.90

75th percentile : 7.20

The top directors

1.First place:

Director: sergio leone

Number of movies: 3

Average score : 8.43

Score standard deviation: 0.45

Percentile rank : 98.7th percentile

2.Director: Christopher Nolan

Number of movies : 8

Average score: 8.43

Score standard deviation: 0.54

Percentile rank : 98.7th percentile

3.Animation Masters:

Director: Pete doctor

Number of movies: 3

Average score : 8.23

Score standard deviation: 0.12

Percentile Rank : 97.7th percentile

Director: Hayao Miyazaki

Number of Movies : 4

Average score : 8.22

Score standard deviation: 0.39

Percentile rank: 97.7th percentile

Insights:

Average score of top directors: 8.04

Average number of movies per top director : 4.4

Score difference from overall mean : 1.58

The analysis shows that the top directors consistently perform well above the overall mean score of 6.47 Both sergio leone and Christopher Nolan lead with identical average scores of 8.43,placing them in the 98.7th percentile.nolan maintained this high average across 8 films,while leone achieved it with 3 classics.

The data also reveals strong showings from animation directors (pete docter,hayao Miyazaki) and auteur filmmakers (quentin tarantino),all scoring above the 97th percentile,demonstrating consistent excellence across multiple films.

E. Budget Analysis:

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

OUTPUT :

The analysis of movie budgets and gross earnings:

Correlation analysis:

Correlation coefficient between budget and gross earnings : 0.0995 indicating a weak positive correlation

This relatively low correlation coefficient suggests that there isn't a strong linear relationship between movie budgets and box office earnings.

Summary statistics (in millions):

Average Budget : $46.24M

Average Gross : $52.61M

Average Profit : $6.38M

Number of movies analyzed : 3756

The movies with the highest profit margins are :

1.James Cameron -

Gross: $760,505,847,

Budget: $237,000,000,

Profit margin : $523,505,847

2. Colin Trevorrow -

Gross: $652,177,271,

Budget: $150,000,000,

Profit Margin: $502,177,271

3.James Cameron -

Gross: $658,672,302,

Budget: $200,000,000,

Profit Margin : $458,672,302

4.George Lucas -

Gross: $460,935,665,

Budget: $11,000,000,

Profit Margin : $449,935,665

5.Steven Spielberg -

Gross: $434,949,459,

Budget: $10,500,000,

Profit Margin: $424,449,459

These movies have the highest profit margins in the dataset.