Name: Naman Rawal

Email: rawalnaman49@gmail.com

IMDB Movie Analysis

Project Description:

The IMDB Movie Analysis project aims to investigate the factors influencing the success of movies based on their IMDB ratings. Understanding what makes a movie successful is crucial for producers, directors, and investors in the film industry. This analysis will provide insights into various elements such as genre, duration, language, director influence, and budget that potentially impact a movie's success.

Approach:

The analysis is conducted in several stages:

1. **Data Cleaning**: The dataset was preprocessed to handle missing values, remove duplicates, and convert data types as necessary. Feature engineering was performed to enhance the dataset for analysis.
2. **Data Analysis**: Various analytical techniques were applied to explore relationships between IMDB ratings and other factors. This included:
   - Analyzing genre distribution and its impact on IMDB scores.
   - Investigating the relationship between movie duration and IMDB scores.
   - Examining the language of the movies and its influence on ratings.
   - Assessing the impact of directors on movie ratings.
   - Exploring the correlation between budgets and financial success.
3. **Five 'Whys' Approach**: This technique was utilized to dive deeper into specific findings and uncover underlying causes behind observed trends.

Tech-Stack Used:

- **Microsoft Excel 2007:** Used for data analysis, including data cleaning, calculations of descriptive statistics, and visualizations such as scatter plots and trendlines.

Insights

A. Movie Genre Analysis

- **Distribution of Genres**: The analysis revealed the most common movie genres within the dataset.
- **Descriptive Statistics**: For each genre, statistics such as mean, median, mode, range, variance, and standard deviation of IMDB scores were calculated.
- **Impact on Ratings**: Insights were drawn regarding how different genres influence the ratings.

Formulas:

To Count: =COUNTIF('cleaned data'!E$2:$E$3849, K2)

Mean: =AVERAGE(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Median: =MEDIAN(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Mode: =MODE(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Max: =MAX(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Min: =MIN(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Variance: =VAR(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Standard Deviation: =STDEV.S(IF('cleaned data'!$E$2:$E$3849=A2, 'cleaned data'!$N$2:$N$3849))

Most Common Genres								
genres	Count	Mean	Median	Mode	Max	Min	Variance	Standard Deviation
Drama	153	7.04183	7.2	7.3	8.8	3.4	0.687054524	0.828887522
Comedy|Drama|Romance	151	6.494702	6.5	6.5	8	4.3	0.562771744	0.750181141
Comedy|Drama	147	6.583673	6.7	6.7	8.8	3.3	0.734800112	0.857204825
Comedy	145	5.84069	6	6.5	8	1.9	1.481874521	1.217322686
Comedy|Romance	135	5.896296	6	6.1	8.4	2.7	0.768269762	0.87650999

B. Movie Duration Analysis

- **Duration Distribution**: Analyzed the distribution of movie durations.
- **Correlation with IMDB Score**: A scatter plot was created to visualize the relationship between movie duration and IMDB score, including a trendline to evaluate the strength of this relationship.
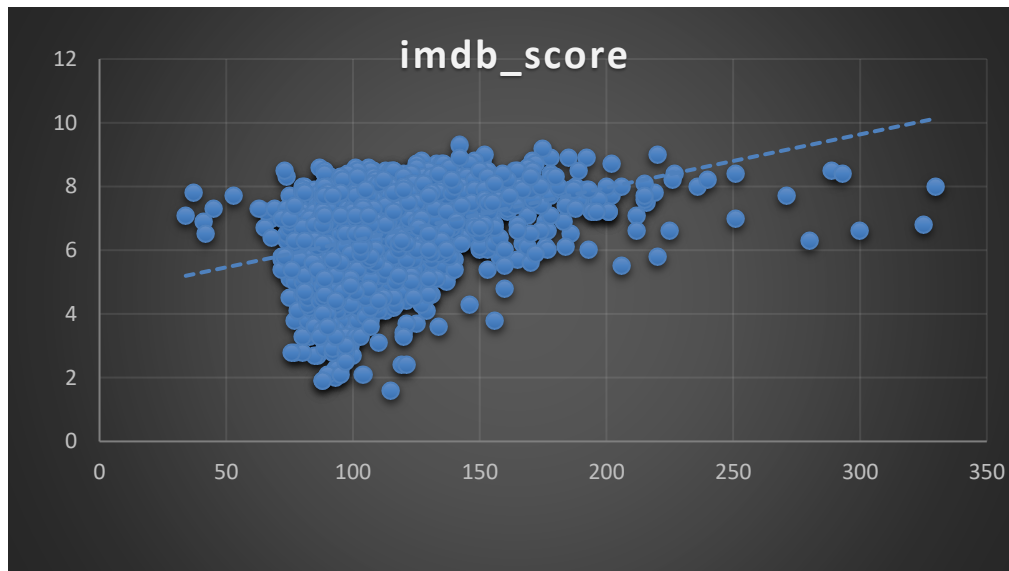
Formulas:

Mean: =AVERAGE(A:A)

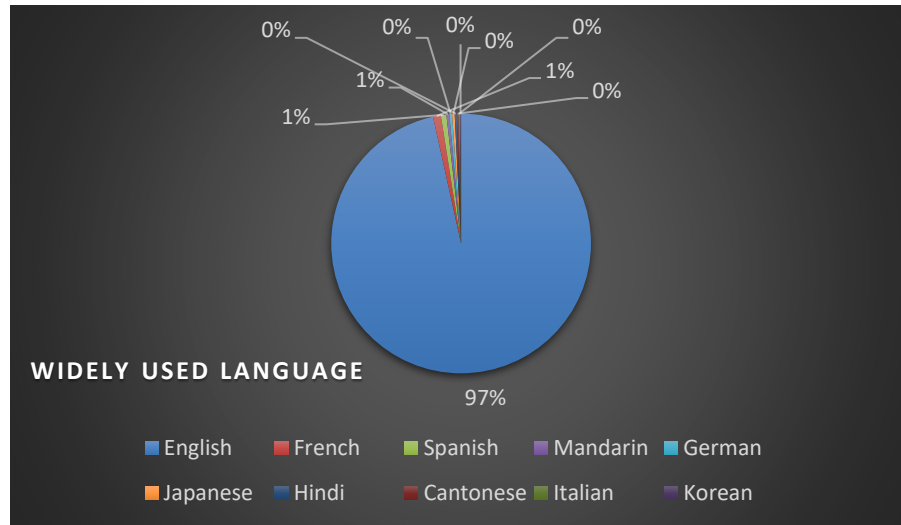**Median:** =MEDIAN(A:A)

Standard deviation: =STDEV.S(A:A)

Average	109.9241164
Median	106
Standard Deviation	22.75364979

C. Language Analysis

- **Common Languages**: The distribution of movies based on language was examined.
- **Statistical Impact**: Descriptive statistics for IMDB scores were calculated for each language to understand its influence on ratings.

Most common Languages				
Language	Count	Mean	Median	Standard Deviation
English	3668	6.423909	6.5	1.048750752
French	37	7.286486	7.2	0.561328861
Spanish	26	7.05	7.15	0.826196103
Mandarin	14	7.021429	7.25	0.765786244
German	13	7.692308	7.7	0.640912811
Japanese	12	7.625	7.8	0.899621132
Hindi	10	6.76	7.05	1.111755369
Cantonese	8	7.2375	7.3	0.440575922
Italian	7	7.185714	7	1.155318962
Korean	5	7.7	7.7	0.570087713

WIDELY USED LANGUAGE

English · French · Spanish · Mandarin · German · Japanese · Hindi · Cantonese · Italian · Korean

D. Director Analysis

- **Top Directors**: Identified directors with the highest average IMDB scores.
- **Success Contribution**: Percentile calculations were used to analyze the contribution of these directors to movie success.

Formulas:

Count: =COUNTIFS('cleaned data'!$J$2:$J$3849, J2)

Mean: =AVERAGE(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

Median: =MEDIAN(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

Standard Deviation: =STDEV.S(IF('cleaned data'!$J$2:$J$3849=J2, 'cleaned data'!$N$2:$N$3849))

| director_name	Average
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

E. **Budget Analysis**

- **Correlation with Financial Success**: The correlation between movie budgets and gross earnings was calculated.
- **Profit Margin Analysis**: The profit margin was assessed, identifying movies with the highest profit margins based on the difference between gross earnings and budget.

CORRELATION
0.100850218

MAX PROFIT	MOVIE TITLE
523505847	AvatarÂ
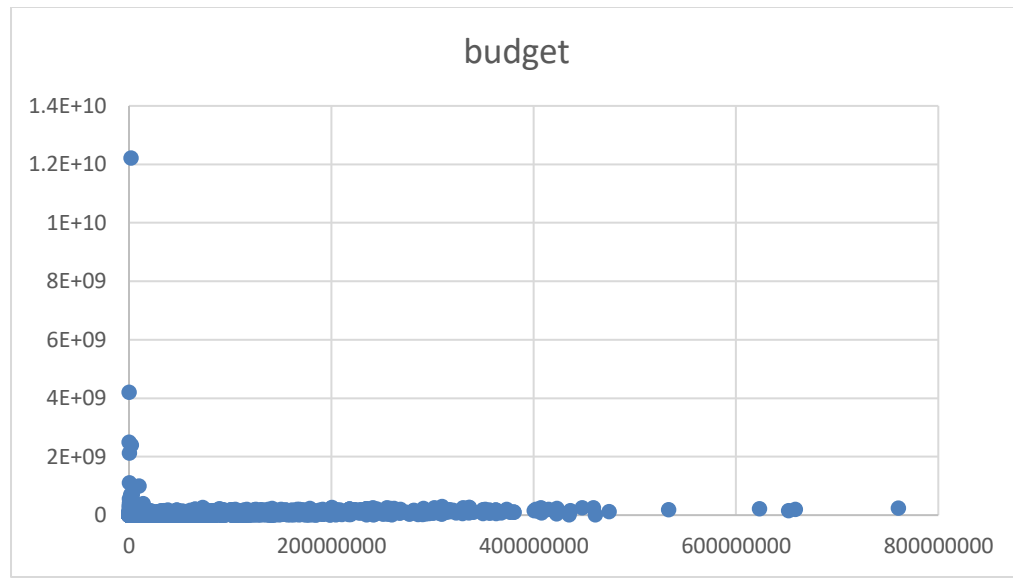
budget

Result:

This project provided a comprehensive analysis of the factors influencing IMDB ratings. Key findings highlight the importance of genre, director reputation, and budget management in determining a movie's success. These insights can guide stakeholders in making informed decisions for future film projects.

Drive Link:

https://docs.google.com/spreadsheets/d/14iUOgu4FLUiWNirB8p-iR3TTICwJsw0a/edit?usp=sharing&ouid=107712337603641298783&rtpof=true&sd=true