**Name: Daksh Rawal**

**Proejct: IMDB Analysis, Final Project 1**

**Project Description:**

The dataset provided is related to IMDB Movies, and the primary problem to investigate is understanding the factors that influence the success of a movie on IMDB. In this context, success is defined by high IMDB ratings. This analysis is of significant value to movie producers, directors, and investors who seek to make informed decisions for their future projects based on what makes a movie successful.

**Tech Stack used:**

Microsoft Excel (2019)

**Project Phases:**

**Data Cleaning:**

In this initial phase, the data will be preprocessed to ensure its suitability for analysis. Data cleaning includes handling missing values, removing duplicates, converting data types if necessary, and potentially performing feature engineering to create new variables that may be useful for the analysis.

**Link for cleaned dataset:**
**https://docs.google.com/spreadsheets/d/1UniAeKK1cXbA8H82Y1OH0ORGehEYsNwZ/edit?usp=drive_link&ouid=102016351939773791513&rtpof=true&sd=true**

**Data Analysis:**

This phase involves exploring the data to gain insights into the relationships between various variables and their impact on IMDB ratings. Factors to be considered include genre, director, budget, year of release, and actors involved. Data analysis will uncover correlations and patterns in the dataset.

**Five 'Whys' Approach:**

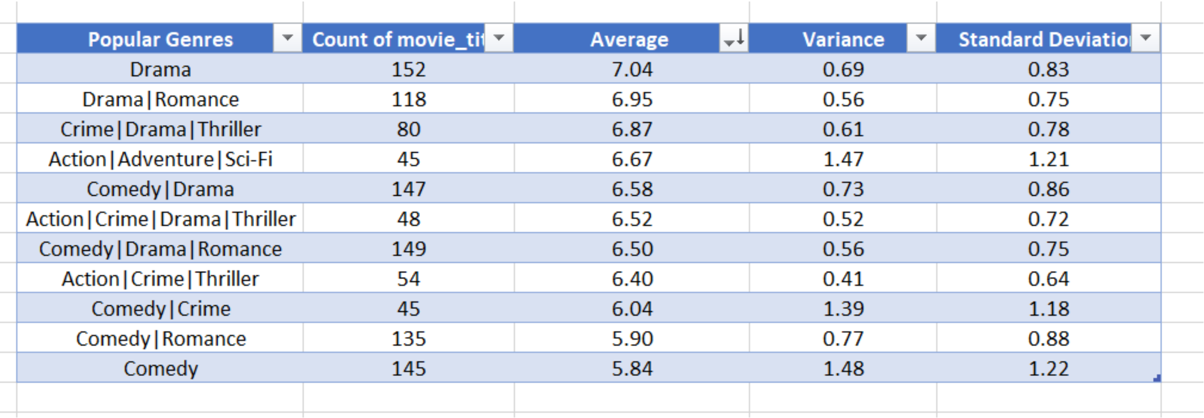
This technique involves delving deeper into the identified correlations and patterns. By asking "Why?" repeatedly, you'll uncover the root causes of these relationships. This understanding will help in providing more meaningful insights that can inform decision-making.

**Report and Data Story:**

After the analysis, a comprehensive report will be created to tell a data-driven story. This report should include the initial problem, key findings, insights gained from the analysis, and actionable recommendations. Visualizations, such as charts and graphs, will be used to make the findings more understandable.

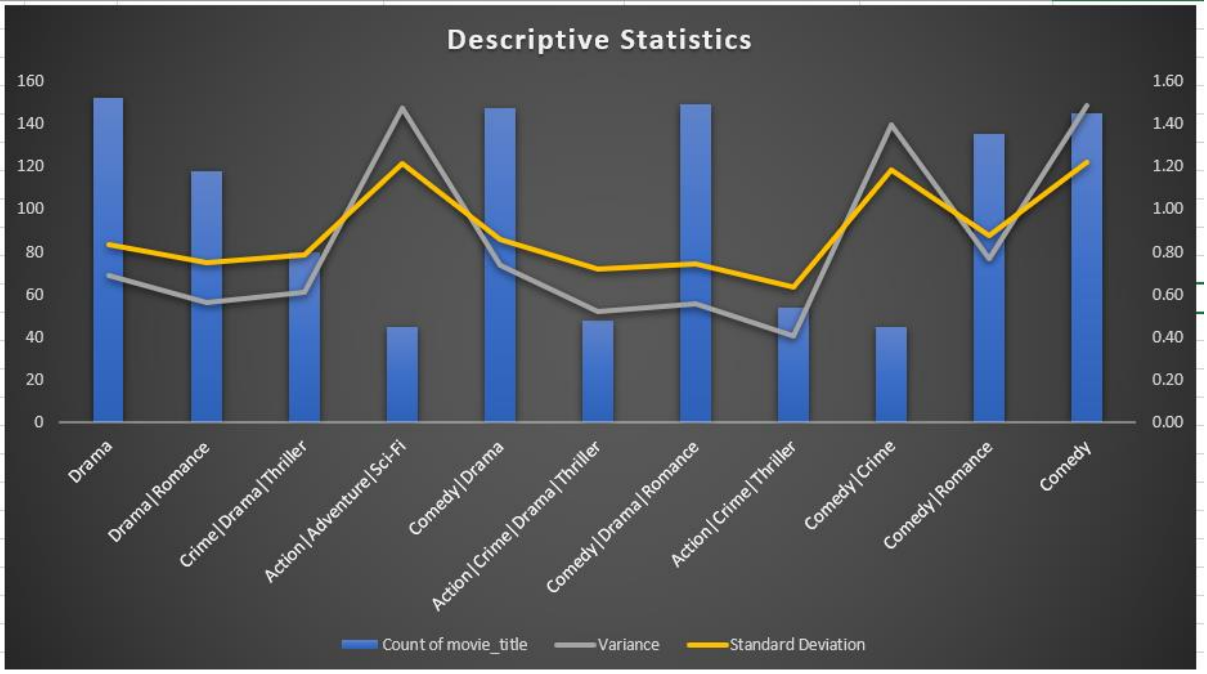
**Results:**

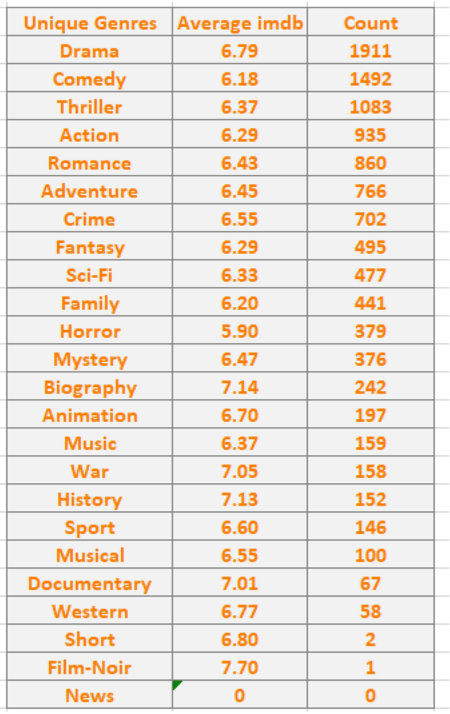
**A. Movie Genre Analysis:**



| Popular Genres | Count of movie_ti | Average | Variance | Standard Deviatio |
|---|---|---|---|---|
| Drama | 152 | 7.04 | 0.69 | 0.83 |
| Drama\|Romance | 118 | 6.95 | 0.56 | 0.75 |
| Crime\|Drama\|Thriller | 80 | 6.87 | 0.61 | 0.78 |
| Action\|Adventure\|Sci-Fi | 45 | 6.67 | 1.47 | 1.21 |
| Comedy\|Drama | 147 | 6.58 | 0.73 | 0.86 |
| Action\|Crime\|Drama\|Thriller | 48 | 6.52 | 0.52 | 0.72 |
| Comedy\|Drama\|Romance | 149 | 6.50 | 0.56 | 0.75 |
| Action\|Crime\|Thriller | 54 | 6.40 | 0.41 | 0.64 |
| Comedy\|Crime | 45 | 6.04 | 1.39 | 1.18 |
| Comedy\|Romance | 135 | 5.90 | 0.77 | 0.88 |
| Comedy | 145 | 5.84 | 1.48 | 1.22 |

Drama is one of the most popular genre that has appeared 152 movies with an average imdb rating of 7.04 followed by Drama|Romance with an average imdb rating of 6.95

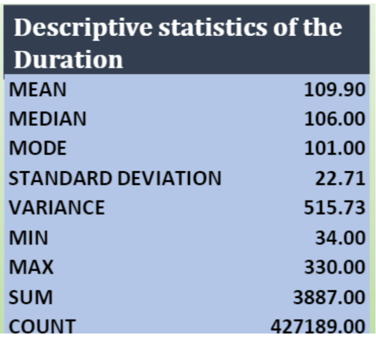




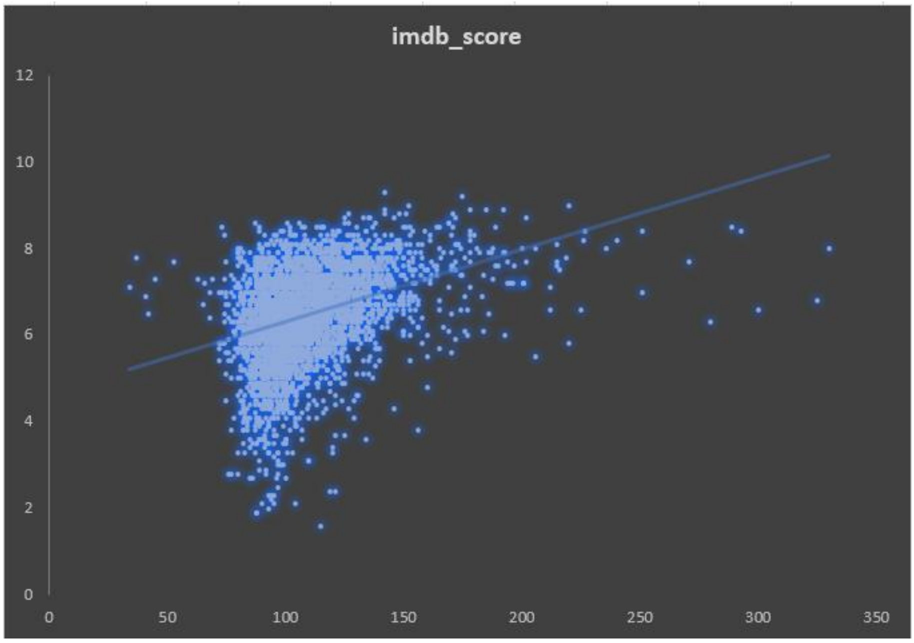
| Unique Genres | Average imdb | Count |
|---|---|---|
| Drama | 6.79 | 1911 |
| Comedy | 6.18 | 1492 |
| Thriller | 6.37 | 1083 |
| Action | 6.29 | 935 |
| Romance | 6.43 | 860 |
| Adventure | 6.45 | 766 |
| Crime | 6.55 | 702 |
| Fantasy | 6.29 | 495 |
| Sci-Fi | 6.33 | 477 |
| Family | 6.20 | 441 |
| Horror | 5.90 | 379 |
| Mystery | 6.47 | 376 |
| Biography | 7.14 | 242 |
| Animation | 6.70 | 197 |
| Music | 6.37 | 159 |
| War | 7.05 | 158 |
| History | 7.13 | 152 |
| Sport | 6.60 | 146 |
| Musical | 6.55 | 100 |
| Documentary | 7.01 | 67 |
| Western | 6.77 | 58 |
| Short | 6.80 | 2 |
| Film-Noir | 7.70 | 1 |
| News | 0 | 0 |

Genres like Action|Adventure|Sci-Fi , Comedy|crime and Comedy have high standard deviation and variance indicating wider range of ratings.

**B. Movie Duration Analysis:**

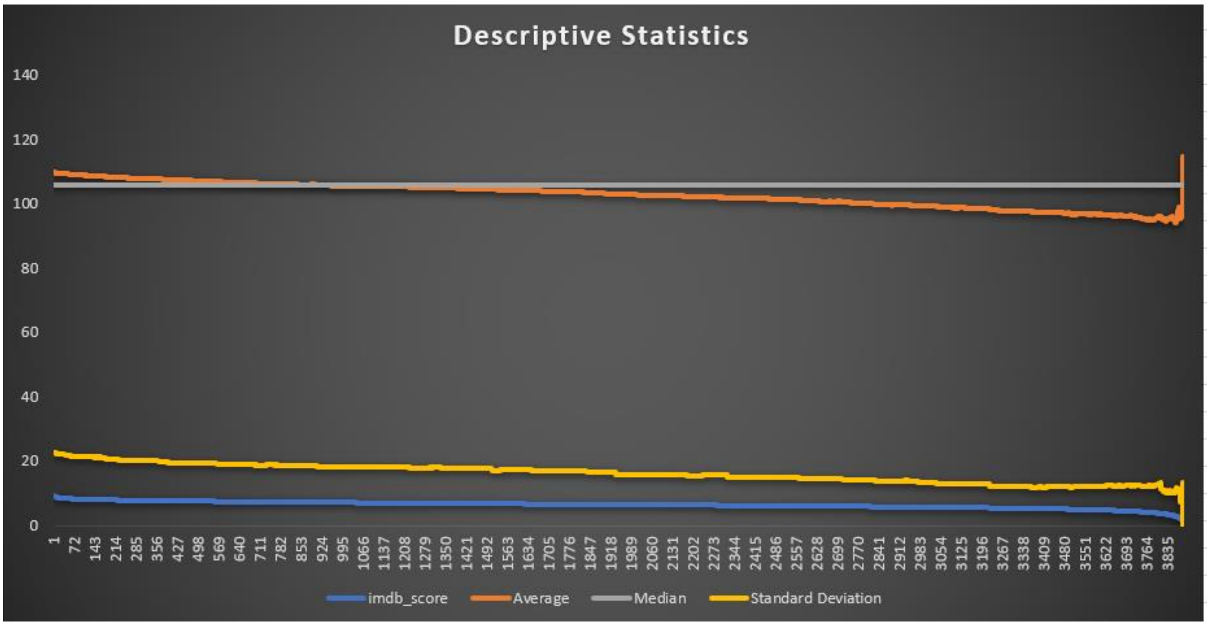


| Descriptive statistics of the Duration | |
|---|---|
| MEAN | 109.90 |
| MEDIAN | 106.00 |
| MODE | 101.00 |
| STANDARD DEVIATION | 22.71 |
| VARIANCE | 515.73 |
| MIN | 34.00 |
| MAX | 330.00 |
| SUM | 3887.00 |
| COUNT | 427189.00 |

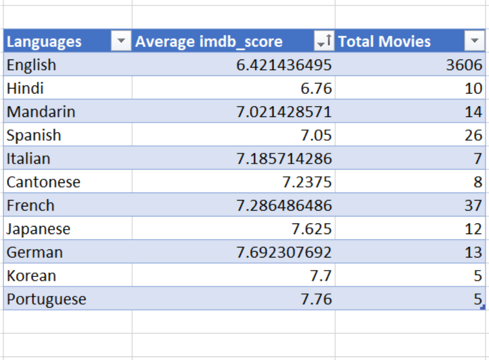


Duration of 80 to 130 has got the maximum of films and the imdb score lies between 4.5 –8.5.

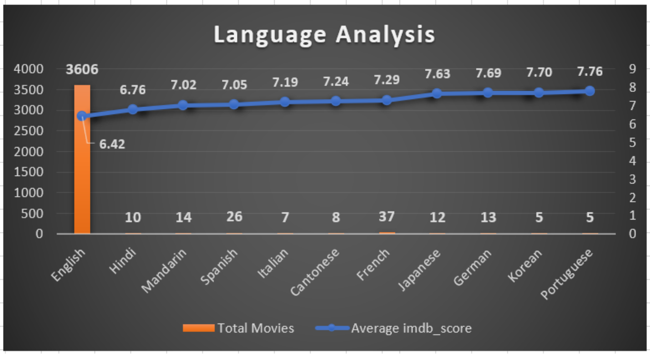
Shortest movie duration is of 34 min whereas longest movie duration is of 330 min



### C. Language Analysis:

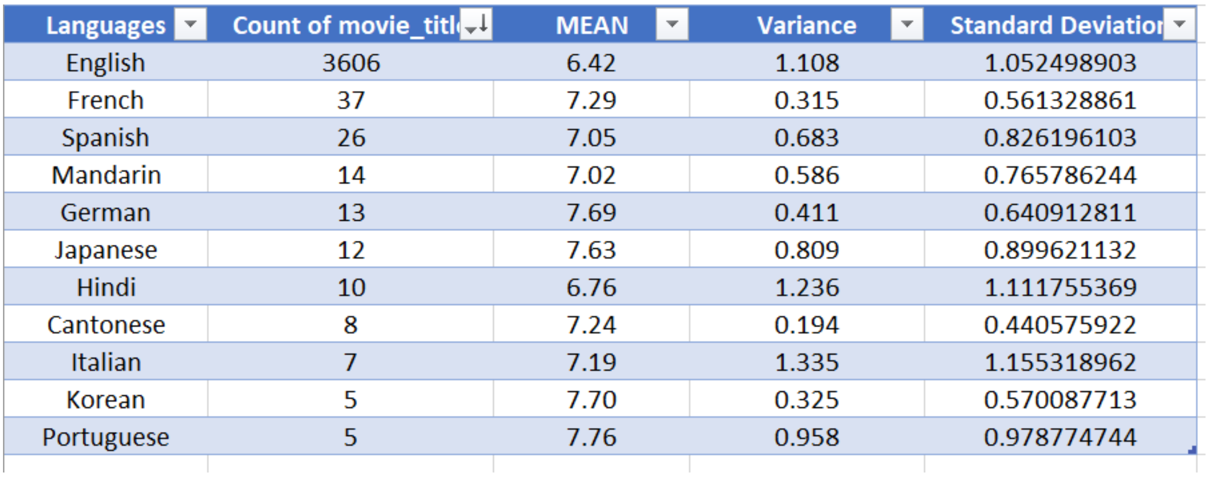


| Languages | Average imdb_score | Total Movies |
|---|---|---|
| English | 6.421436495 | 3606 |
| Hindi | 6.76 | 10 |
| Mandarin | 7.021428571 | 14 |
| Spanish | 7.05 | 26 |
| Italian | 7.185714286 | 7 |
| Cantonese | 7.2375 | 8 |
| French | 7.286486486 | 37 |
| Japanese | 7.625 | 12 |
| German | 7.692307692 | 13 |
| Korean | 7.7 | 5 |
| Portuguese | 7.76 | 5 |



We have used 11 most commonly used languages in movies

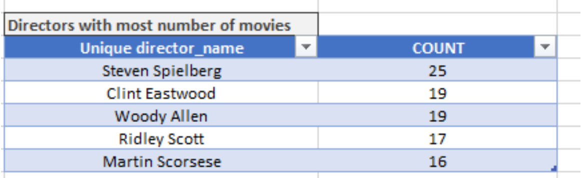
English has the highest number of movies with an average imdb of 6.42Portuguese, Korean, German, Japanese, French, Cantonese, Spanish, Italian, and Mandarin have higher average IMDb ratingsThis is due to consistent audience and fewer movies in these languages



| Languages | Count of movie_titl | MEAN | Variance | Standard Deviatio |
|---|---|---|---|---|
| English | 3606 | 6.42 | 1.108 | 1.052498903 |
| French | 37 | 7.29 | 0.315 | 0.561328861 |
| Spanish | 26 | 7.05 | 0.683 | 0.826196103 |
| Mandarin | 14 | 7.02 | 0.586 | 0.765786244 |
| German | 13 | 7.69 | 0.411 | 0.640912811 |
| Japanese | 12 | 7.63 | 0.809 | 0.899621132 |
| Hindi | 10 | 6.76 | 1.236 | 1.111755369 |
| Cantonese | 8 | 7.24 | 0.194 | 0.440575922 |
| Italian | 7 | 7.19 | 1.335 | 1.155318962 |
| Korean | 5 | 7.70 | 0.325 | 0.570087713 |
| Portuguese | 5 | 7.76 | 0.958 | 0.978774744 |

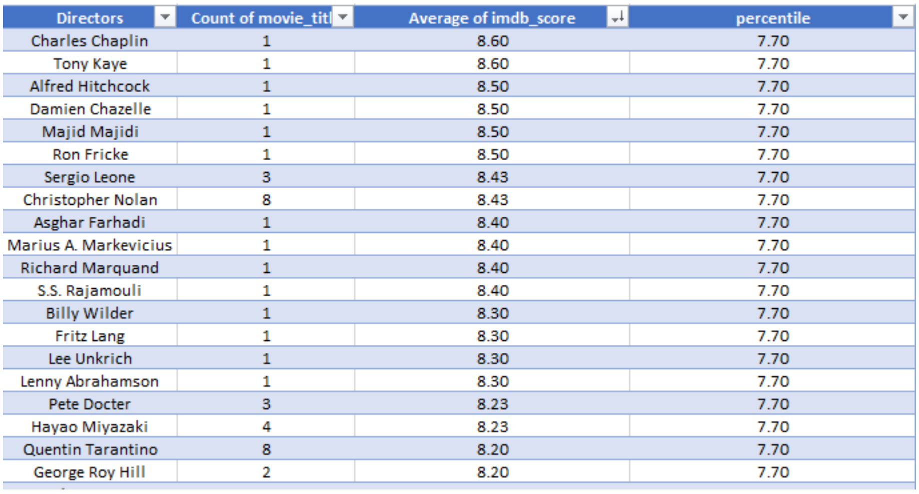
Average movie ratings are consistent across languages ranging from 6.4 –7.7.Higher standard deviation mean more variability whereas variance gives us an idea of how spread the data is across mean

**D. Director Analysis:**



Directors with most number of movies

| Unique director_name | COUNT |
|---|---|
| Steven Spielberg | 25 |
| Clint Eastwood | 19 |
| Woody Allen | 19 |
| Ridley Scott | 17 |
| Martin Scorsese | 16 |



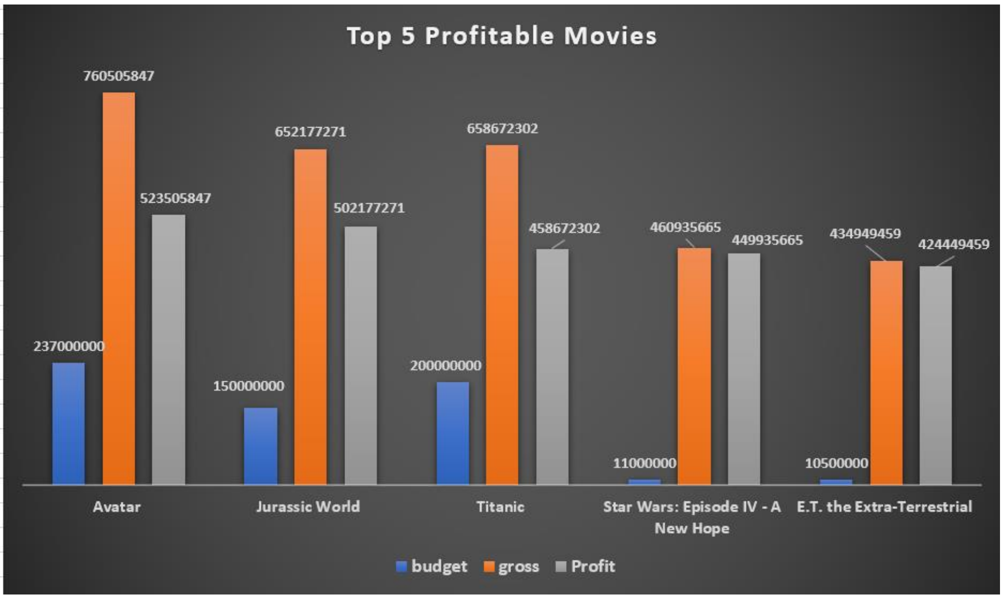
| Directors | Count of movie_titl | Average of imdb_score | percentile |
|---|---|---|---|
| Charles Chaplin | 1 | 8.60 | 7.70 |
| Tony Kaye | 1 | 8.60 | 7.70 |
| Alfred Hitchcock | 1 | 8.50 | 7.70 |
| Damien Chazelle | 1 | 8.50 | 7.70 |
| Majid Majidi | 1 | 8.50 | 7.70 |
| Ron Fricke | 1 | 8.50 | 7.70 |
| Sergio Leone | 3 | 8.43 | 7.70 |
| Christopher Nolan | 8 | 8.43 | 7.70 |
| Asghar Farhadi | 1 | 8.40 | 7.70 |
| Marius A. Markevicius | 1 | 8.40 | 7.70 |
| Richard Marquand | 1 | 8.40 | 7.70 |
| S.S. Rajamouli | 1 | 8.40 | 7.70 |
| Billy Wilder | 1 | 8.30 | 7.70 |
| Fritz Lang | 1 | 8.30 | 7.70 |
| Lee Unkrich | 1 | 8.30 | 7.70 |
| Lenny Abrahamson | 1 | 8.30 | 7.70 |
| Pete Docter | 3 | 8.23 | 7.70 |
| Hayao Miyazaki | 4 | 8.23 | 7.70 |
| Quentin Tarantino | 8 | 8.20 | 7.70 |
| George Roy Hill | 2 | 8.20 | 7.70 |

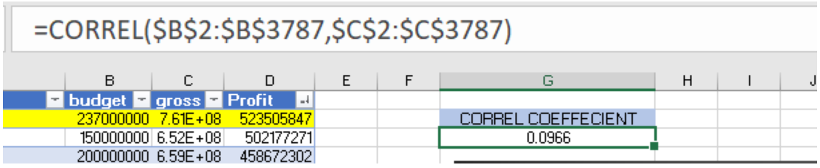
Charles Chaplin and Tony Kaye have the highest average IMDb score of 8.60, with only 1 movie.

Steven Spielberg has the highest average imdb ratings of 7.26 for a total of 25 movies indicating a consistent record

Percentile: Each director's average IMDb score is compared against a common benchmark, to know their relative position in the dataset.

**E. Budget Analysis:**





```
=CORREL($B$2:$B$3787,$C$2:$C$3787)
```

| | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| | budget | gross | Profit | | | | | | |
| | 237000000 | 7.61E+08 | 523505847 | | | CORREL COEFFECIENT | | | |
| | 150000000 | 6.52E+08 | 502177271 | | | 0.0966 | | | |
| | 200000000 | 6.59E+08 | 458672302 | | | | | | |

Correl coefficient indicates the weak correlation between budget and gross across the dataset.Avatar movie has the highest profit.Star wars and Extra terrestrial despite low buget have huge profit.

These data analytics tasks are designed to progressively investigate various aspects of the IMDB movie dataset and reveal meaningful insights. The ultimate goal is to provide actionable recommendations for stakeholders based on the factors that influence a movie's success on IMDB.

**Link for cleaned dataset and ouputs: https://docs.google.com/spreadsheets/d/1UniAeKK1cXbA8H82Y1OH0ORGehEYsNwZ/edit?usp=drive_link&ouid=102016351939773791513&rtpof=true&sd=true**