
IMDB MOVIE
ANALYSIS
Ashiq Paul
Trainity

PROJECT DESCRIPTION

IMDB Movie Analysis

This project involves analyzing an IMDB movie dataset to identify the factors influencing the success of movies, measured by their IMDB ratings. The analysis covers various aspects such as movie genres, durations, languages, directors, budgets, and their relationships to ratings and financial success.

DATASET

Link to the Dataset (Please open on MS Excel)

APPROACH

The Approach taken for this project

Data Cleaning
Cleaned the data by identifying and handling missing values

Exploratory Data Analysis (EDA)
Using Excel and descriptive statistics (mean, median, mode, variance, standard deviation). Pivot tables and formulas like COUNTIF(), AVERAGEIF(), and CORREL() were also used

Visualizations
Bar graphs, scatter plots with trendlines, and pie charts were created to visualize distributions and correlations

Reporting:
Key findings were compiled into a structured report supported by visual evidence, aiming to offer practical recommendations to guide future movies.

TECH STACK

The Tech-Stacks Used

Microsoft Excel 2025

Used for the cleaning of data, pivot table, calculations and chart visualizations

Microsoft Word / Adobe Acrobat / Canva

Microsoft Word Used for preparing and structuring the report and converting to PDF. Canva used for the title page.

Data Cleaning

There are too many columns and there are some which are not required for this analysis, like the color of the movie, imdb link, facebook likes, etc.

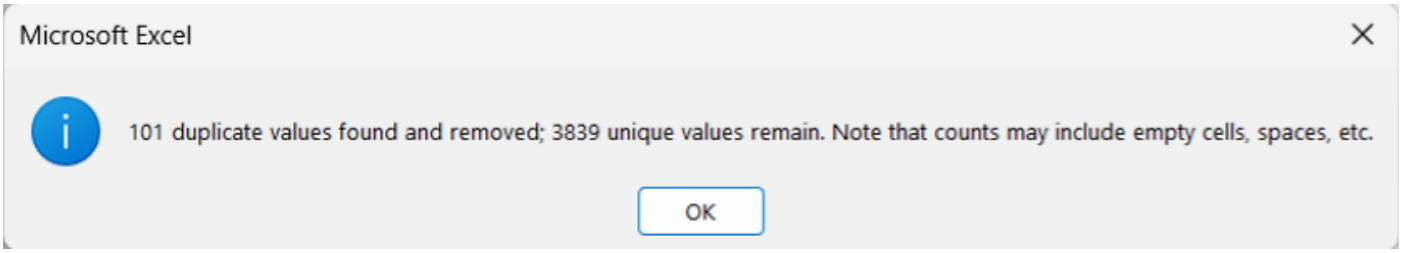
The only columns needed for this analysis are –

- Director_Name
- Duration
- Gross
- Genres
- Movie_Title
- Language
- Country
- Budget
- IMDB_Score

Director_Name	Duration	Gross	Genres	Movie_Title	Language	Country	Budget	IMDB_Score
James Cameron	178	760505847	Action|	AvatarÂ	English	USA	2E+08	7.9
Gore Verbinski	169	309404152	Action|	Pirates of th	English	USA	3E+08	7.1
Sam Mendes	148	200074175	Action|	SpectreÂ	English	UK	2E+08	6.8
Christopher Nola	164	448130642	Action|	The Dark Kn	English	USA	3E+08	8.5
Andrew Stanton	132	73058679	Action|	John Carter	English	USA	3E+08	6.6
Sam Raimi	156	336530303	Action|	Spider-Man	English	USA	3E+08	6.2
Nathan Greno	100	200807262	Adventu	TangledÂ	English	USA	3E+08	7.8
Joss Whedon	141	458991599	Action|	Avengers: A	English	USA	3E+08	7.5
David Yates	153	301956980	Adventu	Harry Potte	English	UK	3E+08	7.5
Zack Snyder	183	330249062	Action|	Batman v Su	English	USA	3E+08	6.9
Bryan Singer	169	200069408	Action|	Superman R	English	USA	2E+08	6.1
Marc Forster	106	168368427	Action|	Quantum of	English	UK	2E+08	6.7
Gore Verbinski	151	423032628	Action|	Pirates of th	English	USA	2E+08	7.3
Gore Verbinski	150	89289910	Action|	The Lone Ra	English	USA	2E+08	6.5
Zack Snyder	143	291021565	Action|	Man of Stee	English	USA	2E+08	7.2
Andrew Adamso	150	141614023	Action|	The Chronic	English	USA	2E+08	6.6
Joss Whedon	173	623279547	Action|	The Avenge	English	USA	2E+08	8.1
Rob Marshall	136	241063875	Action|	Pirates of th	English	USA	3E+08	6.7
Barry Sonnenfelc	106	179020854	Action|	Men in Blac	English	USA	2E+08	6.8
Peter Jackson	164	255108370	Adventu	The Hobbit:	English	New Zea	3E+08	7.5
Marc Webb	153	262030663	Action|	The Amazin	English	USA	2E+08	7
Ridley Scott	156	105219735	Action|	Robin Hood	English	USA	2E+08	6.7
Peter Jackson	186	258355354	Adventu	The Hobbit:	English	USA	2E+08	7.9
Chris Weitz	113	70083519	Adventu	The Golden	English	USA	2E+08	6.1
Peter Jackson	201	218051260	Action|	King KongÂ	English	New Zea	2E+08	7.2
James Cameron	194	658672302	Drama|	TitanicÂ	English	USA	2E+08	7.7

There are blanks present in the dataset. The blanks can either be filled with data from the internet or completely removed. But in this case, there are far too many blanks, so we delete the entire rows with blanks.

There are duplicate rows present as well that needs to be removed.



Some values in the Country column has been replaced after verifying with Google:

- New Line = USA
- Official Site = USA
- West Germany = Germany

A. MOVIE GENRE ANALYSIS

Analyze the distribution of movie genres and their impact on the IMDB score.

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.

The Genre column is split with '|' delimiter and then the number of movies corresponding to that Genre is to be found with the following formula

```
=COUNTIF(IMDB_Movie_Data!D:D,"*" & A2 & "*")
```

('IMDB_Movie_Data' is the name of the data sheet)

Then the Mean is found using the AVERAGEIF() formula

```
=AVERAGEIF(IMDB_Movie_Data!D:D, "*" & A2 & "*", IMDB_Movie_Data!I:I)
```

Then Median is found using the formula

```
=MEDIAN(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

(Since I'm using Excel 2025 I can use the FILTER formula)

Then Mode with the similar formula

```
=MODE(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

Now to find the Range, the difference between the Minimum and Maximum value is found using the formula

```
=MIN(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

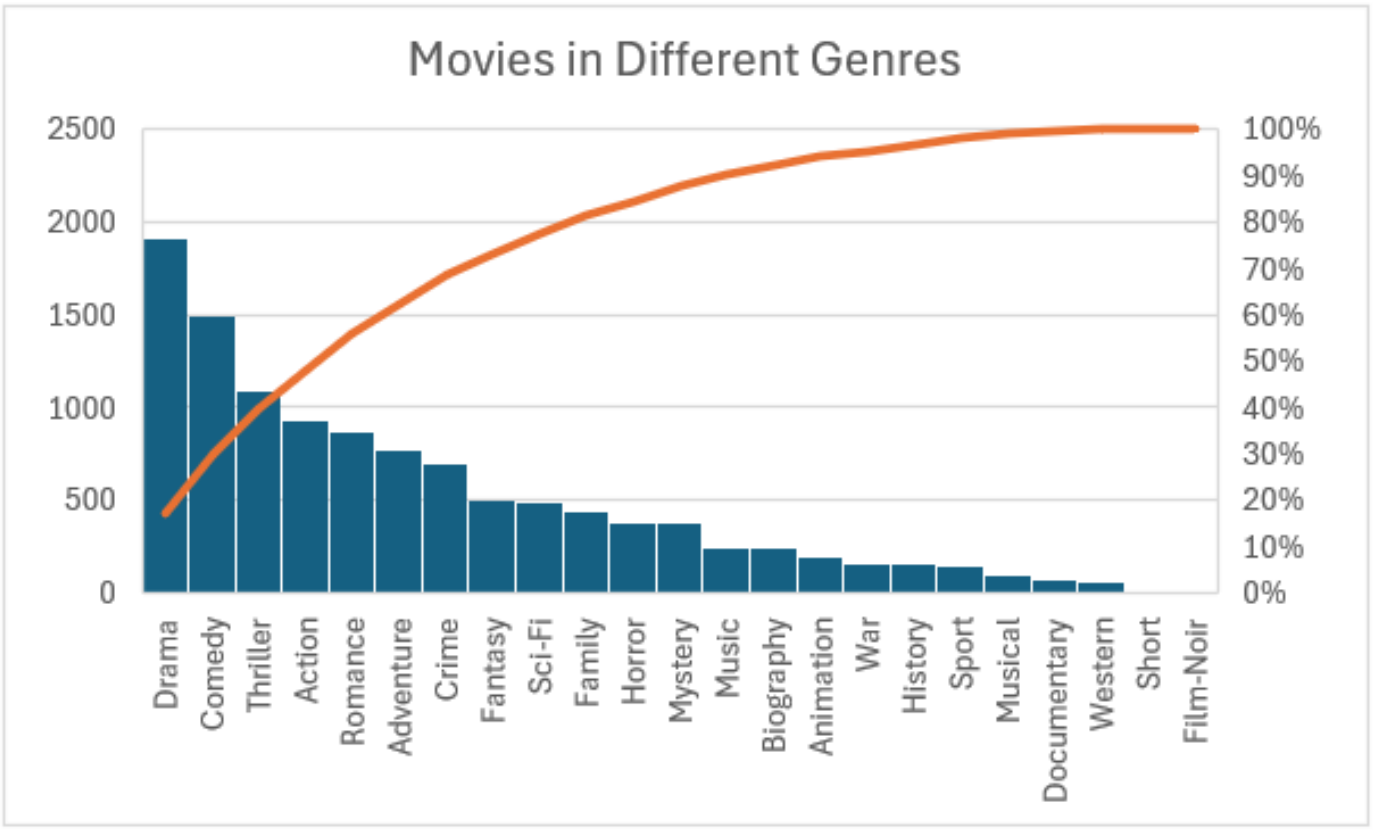
```
=MAX(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

Then finally the Variance and Standard Deviation

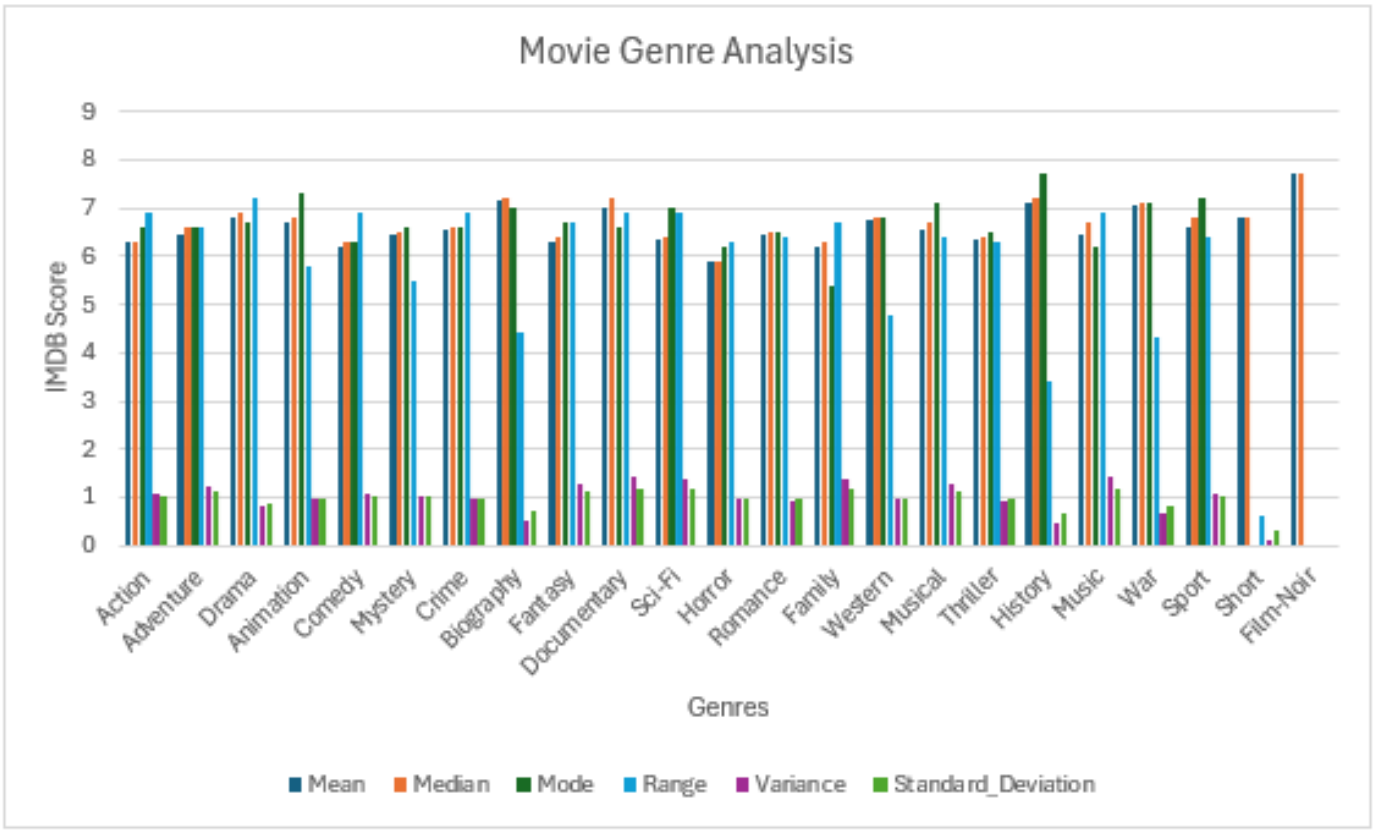
```
=VAR.P(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

```
=STDEV.P(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(A2, IMDB_Movie_Data!D:D))))
```

Genres	Movies	Mean	Median	Mode	Min	Max	Range	Variance	Standard_Deviation
Action	935	6.285989305	6.3	6.6	2.1	9	6.9	1.077033647	1.037802316
Adventure	766	6.454960836	6.6	6.6	2.3	8.9	6.6	1.245895756	1.116197006
Drama	1911	6.789115646	6.9	6.7	2.1	9.3	7.2	0.793581165	0.890831726
Animation	197	6.700507614	6.8	7.3	2.8	8.6	5.8	0.982284006	0.99110242
Comedy	1492	6.183310992	6.3	6.3	1.9	8.8	6.9	1.080706732	1.039570455
Mystery	377	6.469496021	6.5	6.6	3.1	8.6	5.5	1.01214643	1.006054884
Crime	702	6.548148148	6.6	6.6	2.4	9.3	6.9	0.967083465	0.983404019
Biography	242	7.140082645	7.2	7	4.5	8.9	4.4	0.502153712	0.708628049
Fantasy	496	6.285080645	6.4	6.7	2.2	8.9	6.7	1.297922574	1.139264049
Documentary	67	7.011940299	7.2	6.6	1.6	8.5	6.9	1.418364892	1.190951255
Sci-Fi	484	6.327272727	6.4	7	1.9	8.8	6.9	1.359504132	1.165977758
Horror	379	5.903957784	5.9	6.2	2.3	8.6	6.3	0.979535787	0.989715003
Romance	866	6.426212471	6.5	6.5	2.1	8.5	6.4	0.937869488	0.968436621
Family	441	6.2	6.3	5.4	1.9	8.6	6.7	1.364807256	1.168249655
Western	58	6.765517241	6.8	6.8	4.1	8.9	4.8	0.979845422	0.989871417
Musical	102	6.550980392	6.7	7.1	2.1	8.5	6.4	1.29485198	1.13791563
Thriller	1087	6.372309108	6.4	6.5	2.7	9	6.3	0.938248854	0.968632466
History	152	7.131578947	7.2	7.7	5.5	8.9	3.4	0.448608033	0.669782079
Music	247	6.456680162	6.7	6.2	1.6	8.5	6.9	1.407637562	1.186439026
War	159	7.048427673	7.1	7.1	4.3	8.6	4.3	0.648283691	0.805160662
Sport	147	6.601360544	6.8	7.2	2	8.4	6.4	1.091290666	1.044648585
Short	2	6.8	6.8	0	6.5	7.1	0.6	0.09	0.3
Film-Noir	1	7.7	7.7	0	7.7	7.7	0	0	0



The top 5 Genres are Drama, Comedy, Thriller, Action and Romance



The descriptive statistics of the top 5 genres are almost at the highest level

B. MOVIE DURATION ANALYSIS

Analyze the distribution of movie durations and its impact on the IMDB score.

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

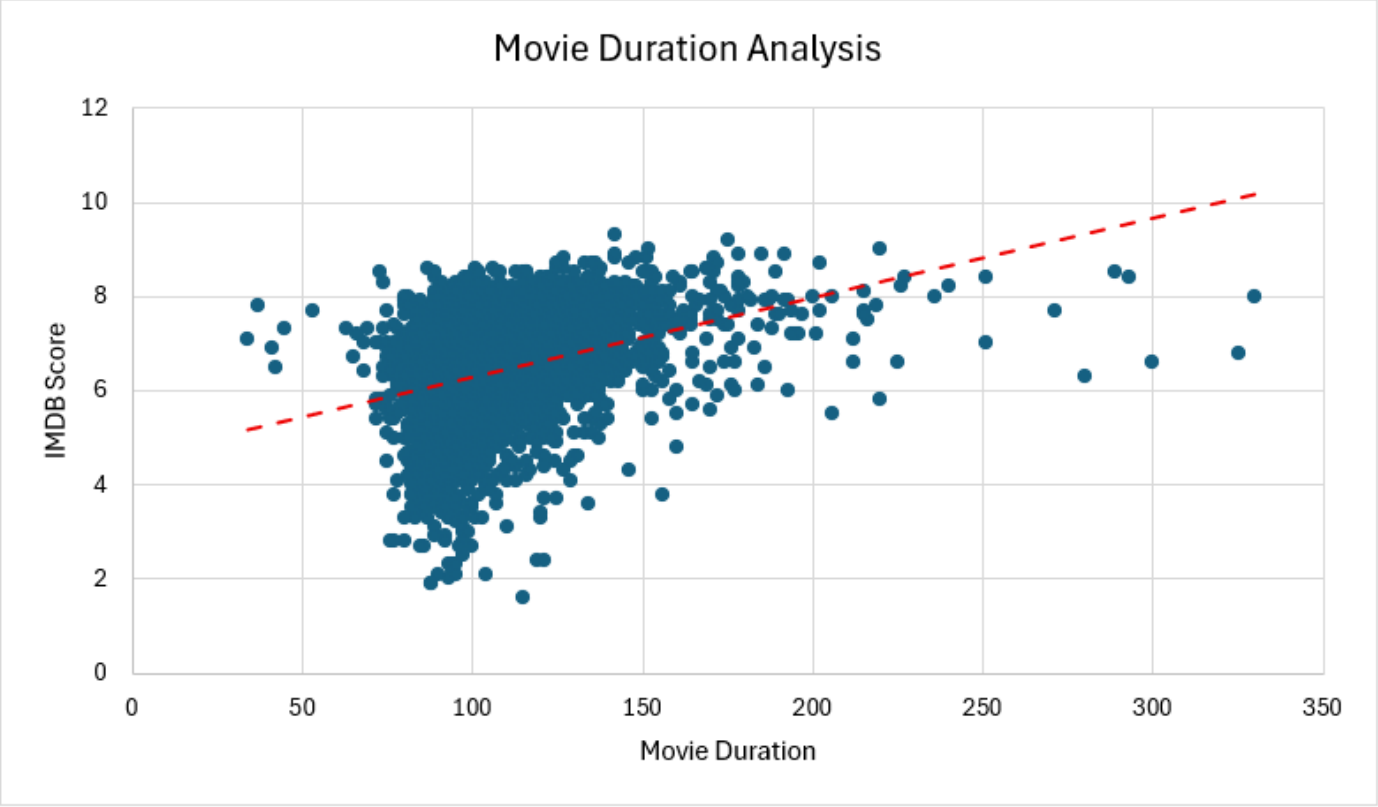
The Mean, Median and Standard Deviation of the duration table is found

```
=AVERAGE(IMDB_Movie_Data!B:B)
```

```
=MEDIAN(IMDB_Movie_Data!B:B)
```

```
=STDEV.P(IMDB_Movie_Data!B:B)
```

	Mean	Median	Standard_Deviation
Movie Duration	109.808505	105	22.76019457



This scatter plot shows that the Movie Duration and IMDB Score have a positive relationship

C. LANGUAGE ANALYSIS

Examine the distribution of movies based on their language.

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

The number of movies of that specific language, Mean, Median and Standard Deviation is found

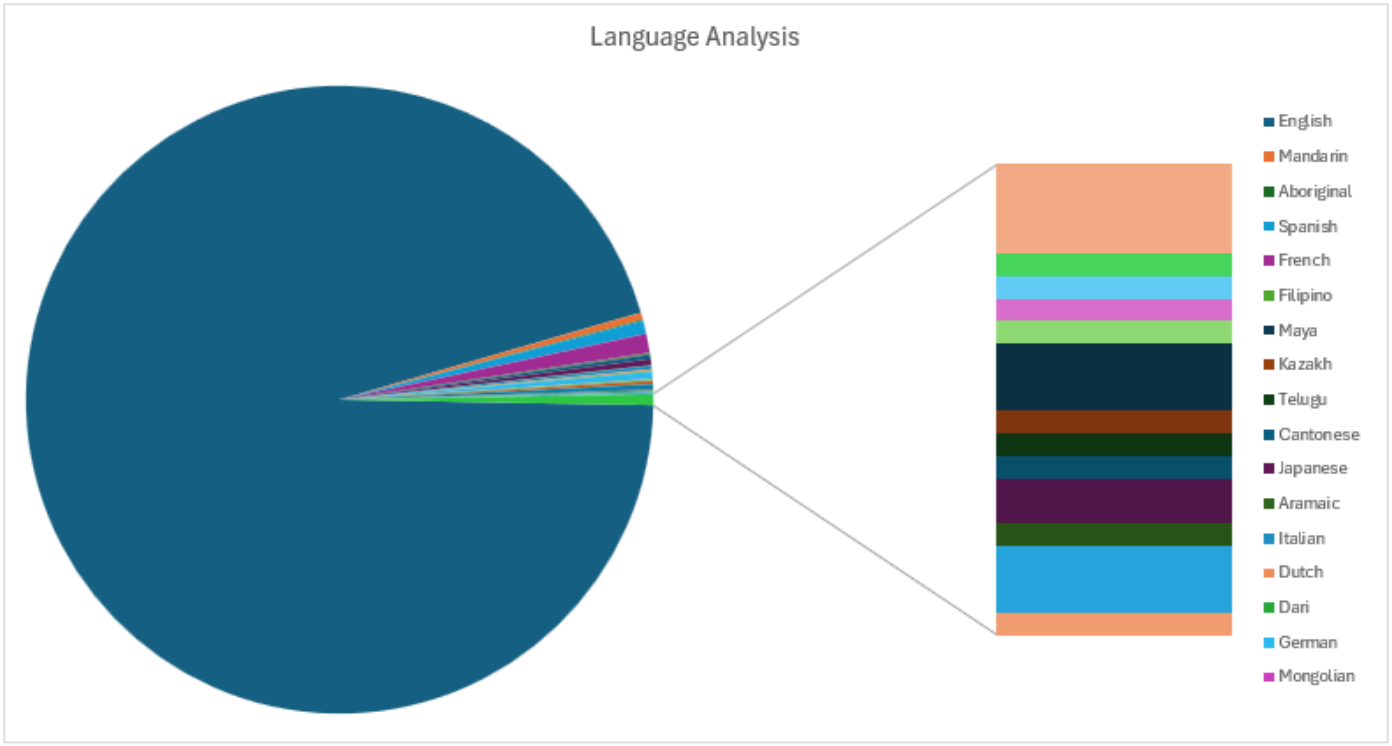
```
=COUNTIF(IMDB_Movie_Data!F:F,D5)
```

```
=AVERAGEIF(IMDB_Movie_Data!F:F,D5,IMDB_Movie_Data!I:I)
```

```
=MEDIAN(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(D5, IMDB_Movie_Data!F:F))))
```

```
=STDEV.P(FILTER(IMDB_Movie_Data!I:I, ISNUMBER(SEARCH(D5, IMDB_Movie_Data!F:F))))
```

Languages	Movies	Mean	Median	Standard Deviation
English	3606	6.42143649	6.5	1.052352956
Mandarin	14	7.02142857	7.25	0.737930089
Aboriginal	2	6.95	6.95	0.55
Spanish	26	7.05	7.15	0.810151933
French	37	7.28648649	7.2	0.553691378
Filipino	1	6.7	6.7	0
Maya	1	7.8	7.8	0
Kazakh	1	6	6	0
Telugu	1	8.4	8.4	0
Cantonese	8	7.2375	7.3	0.412121038
Japanese	12	7.625	7.8	0.861321659
Aramaic	1	7.1	7.1	0
Italian	7	7.18571429	7	1.069617517
Dutch	3	7.56666667	7.8	0.329983165
Dari	2	7.5	7.4	0.709065186
German	13	7.69230769	7.7	0.615769111
Mongolian	1	7.3	7.3	0
Thai	3	6.63333333	6.6	0.368178701
Bosnian	1	4.3	4.3	0
Korean	5	7.7	7.7	0.509901951
Hungarian	1	7.1	7.1	0
Hindi	10	6.76	7.05	1.05470375
Icelandic	1	6.9	6.9	0
Danish	3	7.9	8.1	0.43204938
Portuguese	5	7.76	8	0.875442745
Norwegian	4	7.15	7.3	0.497493719
Czech	1	7.4	7.4	0
Russian	1	6.5	6.5	0
None	1	8.5	8.5	0
Zulu	1	7.3	7.3	0
Hebrew	3	7.5	7.3	0.355902608
Dzongkha	1	7.5	7.5	0
Arabic	1	7.2	7.2	0
Vietnamese	1	7.4	7.4	0
Indonesian	2	7.9	7.9	0.3
Romanian	1	7.9	7.9	0
Persian	3	8.13333333	8.4	0.449691252
Swedish	1	7.6	7.6	0



English is the most common language of the movies then comes French and Spanish

D. DIRECTOR ANALYSIS

Influence of directors on movie ratings.

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

First the Director_Name column is copied and then the duplicates were removed. And then the number of movies corresponding to that Director is found using

```
=COUNTIF(IMDB_Movie_Data!A:A, D5)
```

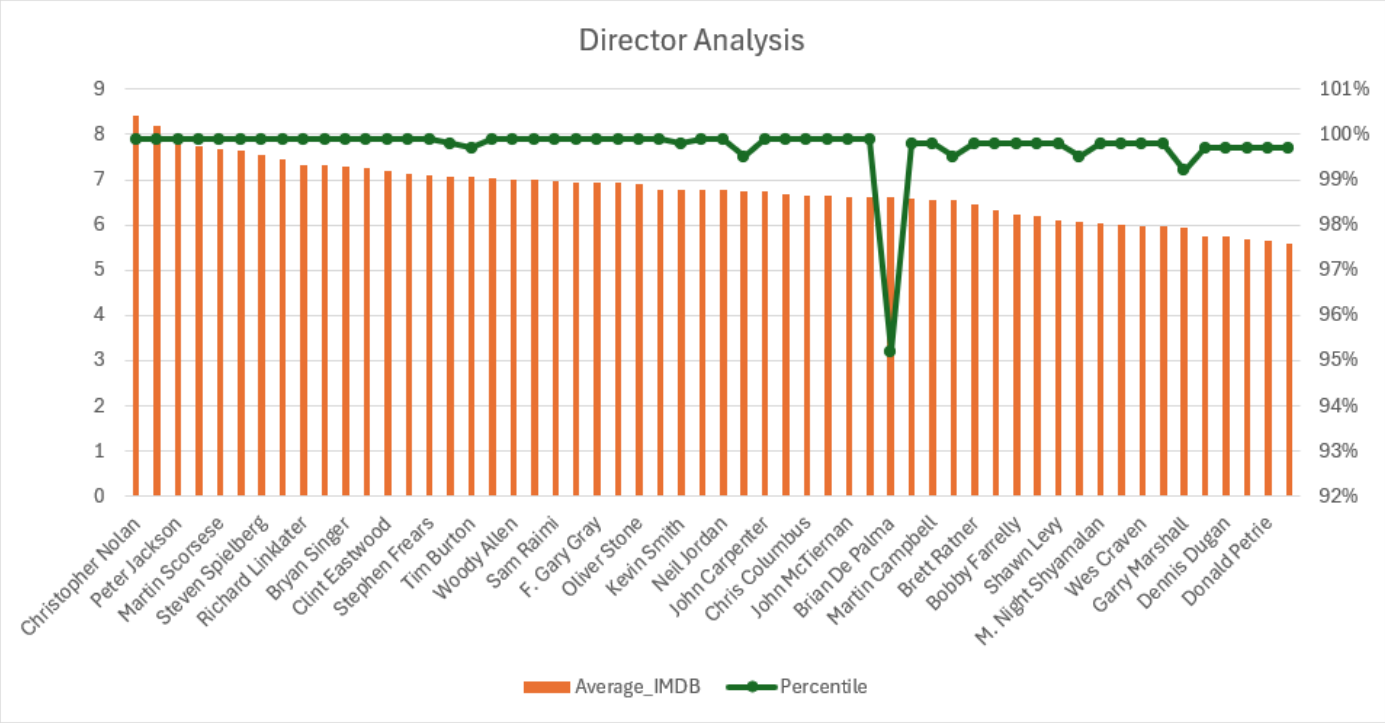
Then the Average IMDB Score for the movies of that Director is found using

```
=AVERAGEIF(IMDB_Movie_Data!A:A, D5, IMDB_Movie_Data!I:I)
```

Finally the Percentile is found using

```
=PERCENTRANK.EXC(F5:F1755, F5)
```

Director_Name	Movies	Average_IMDB	Percentile
Christopher Nolan	8	8.425	0.999
Quentin Tarantino	8	8.2	0.999
Peter Jackson	9	7.888888889	0.999
David Fincher	10	7.75	0.999
Martin Scorsese	16	7.675	0.999
Francis Ford Coppola	9	7.655555556	0.999
Steven Spielberg	25	7.544	0.999
Danny Boyle	8	7.4375	0.999
Richard Linklater	11	7.327272727	0.999
Robert Zemeckis	13	7.307692308	0.999
Bryan Singer	8	7.2875	0.999
Ang Lee	8	7.25	0.999
Clint Eastwood	19	7.205263158	0.999
Ridley Scott	16	7.13125	0.999
Stephen Frears	8	7.0875	0.999
James Mangold	8	7.075	0.998
Tim Burton	14	7.05	0.997
Rob Reiner	11	7.018181818	0.999
Woody Allen	19	7	0.999
Lasse HallstrÃ¶m	8	6.9875	0.999
Sam Raimi	10	6.96	0.999
Antoine Fuqua	8	6.9375	0.999
F. Gary Gray	8	6.9375	0.999
Ron Howard	13	6.930769231	0.999
Oliver Stone	13	6.907692308	0.999
Tony Scott	12	6.791666667	0.999
Kevin Smith	10	6.78	0.998
Phillip Noyce	9	6.766666667	0.999
Neil Jordan	8	6.7625	0.999
Spike Lee	15	6.733333333	0.995
John Carpenter	10	6.73	0.999
Steven Soderbergh	15	6.68	0.999
Chris Columbus	11	6.654545455	0.999
Richard Donner	9	6.633333333	0.999
John McTiernan	10	6.63	0.999
Michael Bay	12	6.616666667	0.999
Brian De Palma	10	6.6	0.952
Barry Levinson	13	6.576923077	0.998
Martin Campbell	8	6.55	0.998
Harold Ramis	8	6.55	0.995
Brett Ratner	9	6.455555556	0.998
Joel Schumacher	12	6.341666667	0.998
Bobby Farrelly	9	6.244444444	0.998
Roland Emmerich	8	6.1875	0.998
Shawn Levy	11	6.090909091	0.998
Ivan Reitman	8	6.075	0.995
M. Night Shyamalan	9	6.044444444	0.998
Paul W.S. Anderson	10	5.99	0.998
Wes Craven	10	5.97	0.998
Adam Shankman	8	5.9625	0.998
Garry Marshall	8	5.95	0.992
Renny Harlin	15	5.746666667	0.997
Dennis Dugan	9	5.733333333	0.997
Robert Rodriguez	13	5.692307692	0.997
Donald Petrie	8	5.6625	0.997
Rob Cohen	9	5.577777778	0.997



Only the Directors with movie count 8 or above are shown, otherwise it would be unfair to compare them with Directors with few good movies.

E. BUDGET ANALYSIS

Explore the relationship between movie budgets and their financial success.

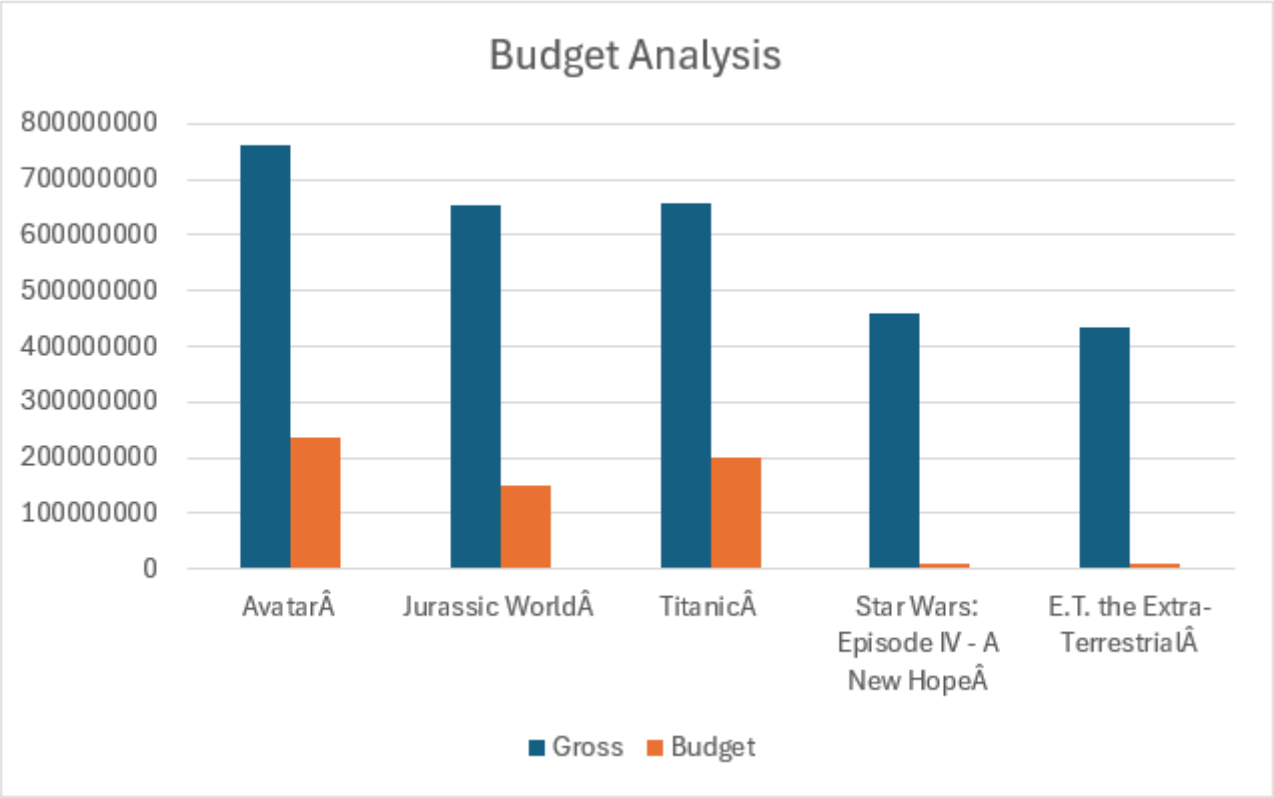
Analyze the correlation between movie budgets and gross earnings, and identify the movies with the highest profit margin.

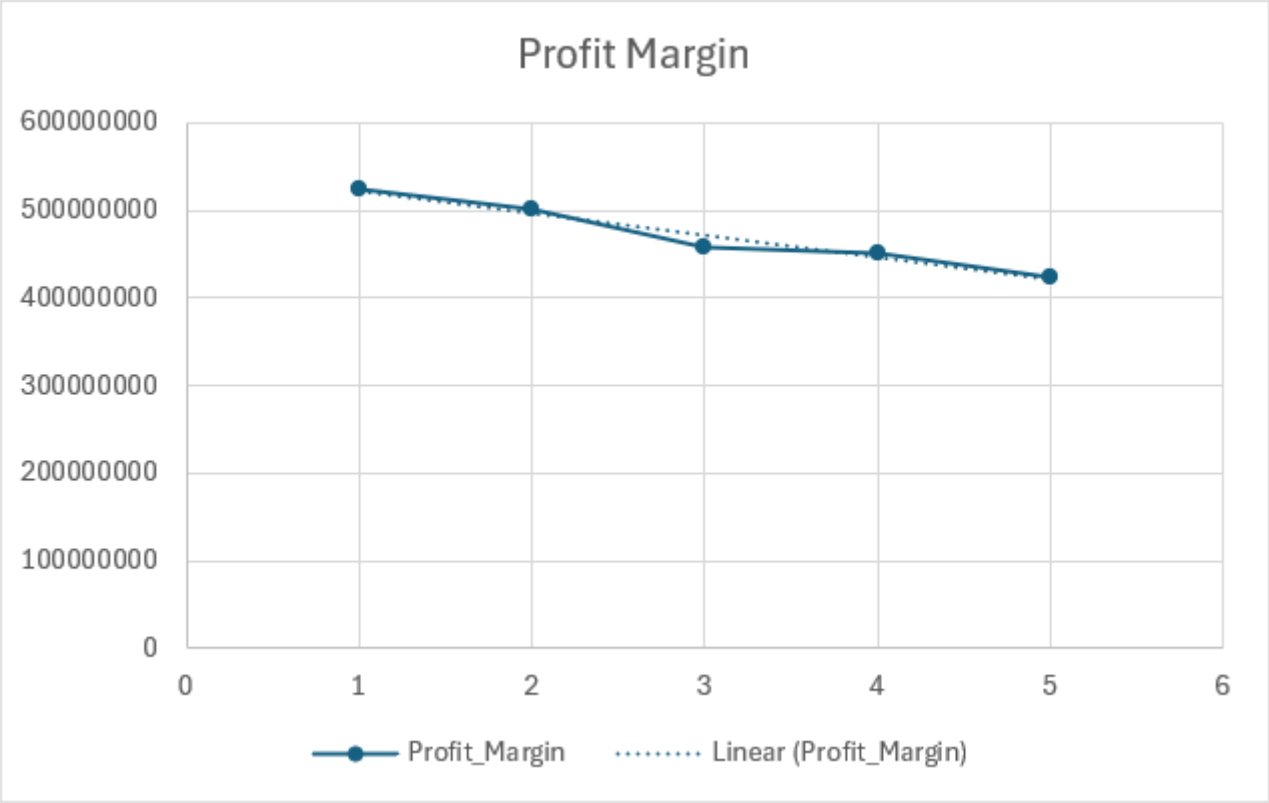
Movie_Title	Gross	Budget	Profit_Margin
AvatarÂ	760505847	237000000	523505847
Jurassic WorldÂ	652177271	150000000	502177271
TitanicÂ	658672302	200000000	458672302
Star Wars: Episode IV - A New HopeÂ	460935665	11000000	449935665
E.T. the Extra-TerrestrialÂ	434949459	10500000	424449459

These are the top 5 movies with highest profit margin, and the movie Avatar is the movie with the highest profit margin.

The Correlation between Gross and Budget is

Correlation	0.096569





The movies with higher budget tends have a higher chance to produce higher profit.

RESULT

This project helped me strengthen my skills in data cleaning, analysis, and statistics. By exploring factors like genre, duration, language, directors, and budget, I was able to identify patterns that influence a movie's success on IMDB. I learned that while high budgets and popular genres can play a role, strong direction often have a greater impact on ratings. Overall, this analysis gave me valuable insights into how data can guide creative and business decisions in the film industry.

THANK YOU
