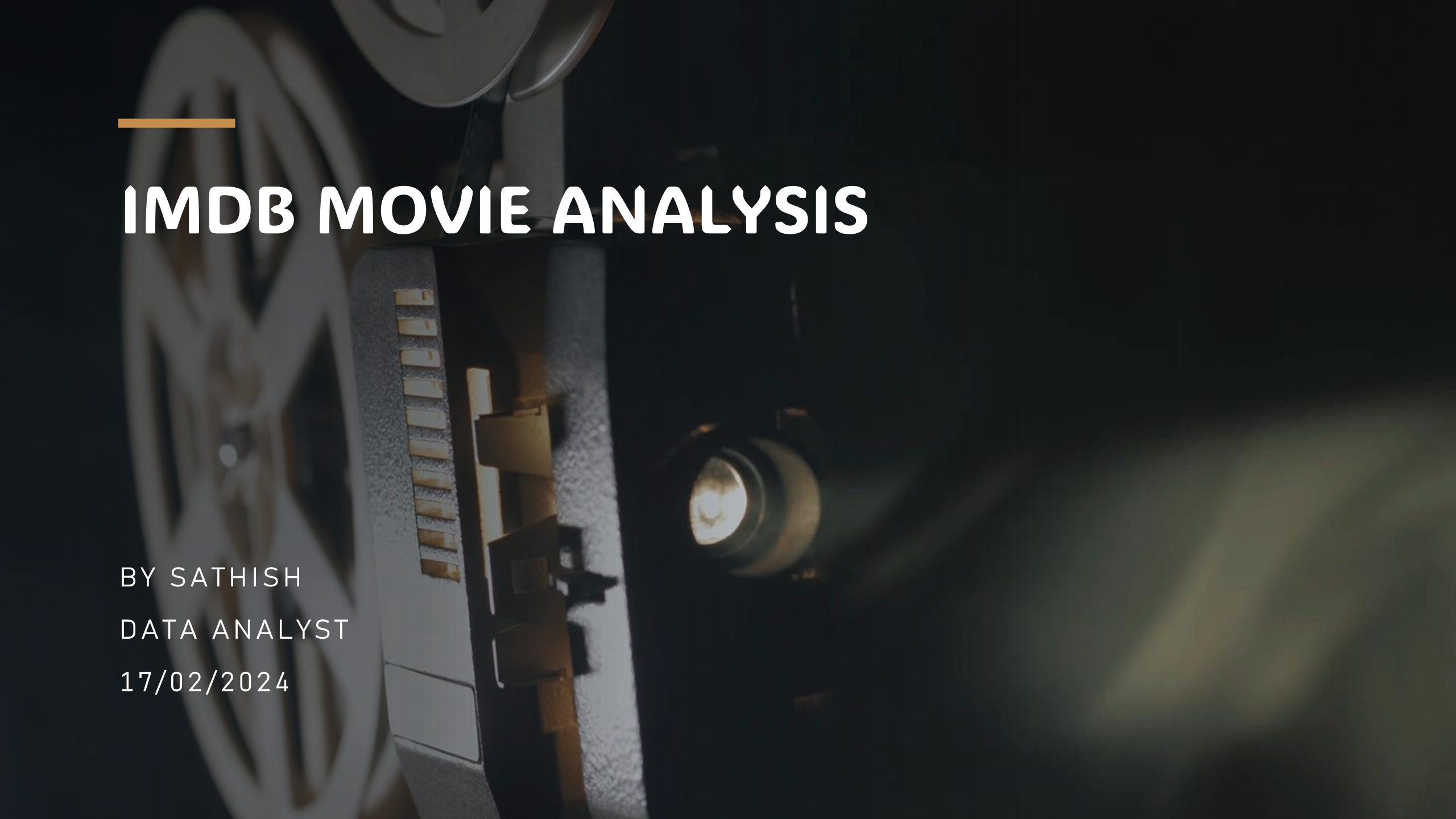


IMDB MOVIE ANALYSIS

BY SATHISH

DATA ANALYST

17/02/2024

Project Description

The objective of this project is to investigate and understand the reasons behind the success of movies on IMDB.

Approach

- Five why’s approach.
- <u>Data cleaning and wrangling</u>

1. Identifying and Removing Irrelevant Columns (From the perspective of the planned analysis)
2. Removed duplicate records.
3. Removed rows with potential null values with the view to maintain data quality and accuracy.
4. Calculated the profit for each movie by subtracting gross collection from the budget.

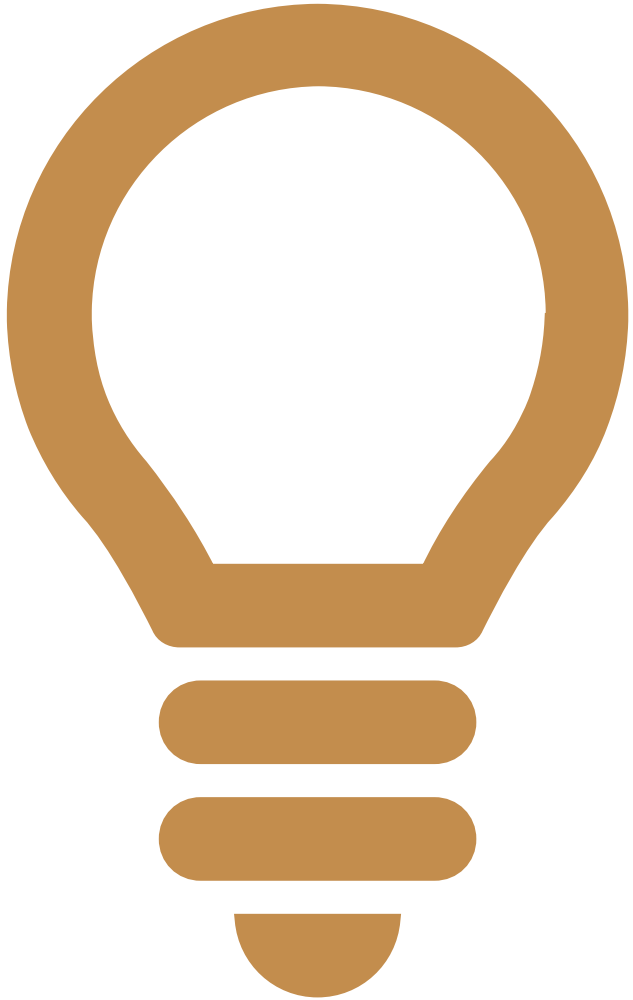
- <u>Assumption</u>: Assumed columns that represents the currency are in terms of Rupees.

Tech Stacks Used

All the analysis process including data cleaning, manipulation, statistical analysis and visualizations are done using Microsoft Excel 365. Since dataset was small, and is in requirement of use of statistical analysis, I preferred Excel.

Microsoft PowerPoint – For the communication of the insights derived to the stake holders.

Insights



Movie Genre Analysis

Task: To determine most common genres of movies in the dataset and to calculate the descriptive statistics for the same.

Workings

unique_genre	freq.	mean_imdb	median_imdb	mode_imdb	max_imdb	min_imdb	range_imdb	var_imdb	stdv_imdb
Drama	1941	6.79	6.90	6.70	9.30	2.10	7.20	0.79	0.89
Comedy	1504	6.18	6.30	6.70	8.80	1.90	6.90	1.08	1.04
Thriller	1117	6.38	6.40	6.50	9.00	2.70	6.30	0.93	0.97
Action	962	6.29	6.30	6.10	9.00	2.10	6.90	1.06	1.03
Romance	878	6.43	6.50	6.50	8.50	2.10	6.40	0.93	0.97
Adventure	787	6.46	6.60	6.70	8.90	2.30	6.60	1.23	1.11
Crime	714	6.54	6.60	6.60	9.30	2.40	6.90	0.96	0.98
Fantasy	514	6.29	6.40	6.70	8.90	2.20	6.70	1.28	1.13
Sci-Fi	497	6.32	6.40	6.70	8.80	1.90	6.90	1.34	1.16
Family	450	6.21	6.30	6.70	8.60	1.90	6.70	1.35	1.16
Horror	391	5.93	6.00	5.90	8.60	2.30	6.30	0.99	1.00
Mystery	383	6.48	6.50	6.60	8.60	3.10	5.50	1.01	1.01
Biography	243	7.14	7.20	7.00	8.90	4.50	4.40	0.50	0.71
Animation	199	6.70	6.80	6.70	8.60	2.80	5.80	0.98	0.99
War	160	7.05	7.10	7.10	8.60	4.30	4.30	0.65	0.81
Music	248	6.46	6.70	6.20	8.50	1.60	6.90	1.41	1.19
History	153	7.14	7.20	7.70	8.90	5.50	3.40	0.45	0.67
Sport	151	6.60	6.80	7.20	8.40	2.00	6.40	1.09	1.04
Musical	103	6.56	6.70	7.10	8.50	2.10	6.40	1.30	1.14
Documentary	64	6.99	7.20	6.60	8.50	1.60	6.90	1.50	1.22
Western	58	6.77	6.80	6.80	8.90	4.10	4.80	1.00	1.00
Short	2	6.80	6.80	NA	7.10	6.50	0.60	0.18	0.42
Film-Noir	1	7.70	7.70	NA	7.70	7.70	0.00	NA	NA

Insights

- If we go with the frequency, "Drama" is the most common genre with 1941 movies listed. The mean IMDB for the same is 6.79.
- "Comedy" and "Thriller" are other common genres with 1504 and 1117 movies listed and with mean IMDB score of 6.18 and 6.38 respectively.
- "Short" and "Film-Noir" are the least common genre with only 2 and 1 movies listed. The mean IMDB scores are 6.8 and 7.7 respectively.
- The mean IMDB score of "Film-Noir" genre i.e., 7.7, is the highest mean IMDB score among others with only 1 movie listed.
- "Horror" genre, with 391 movies listed, has the least mean IMDB score of 5.93.

Movie Duration Analysis

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

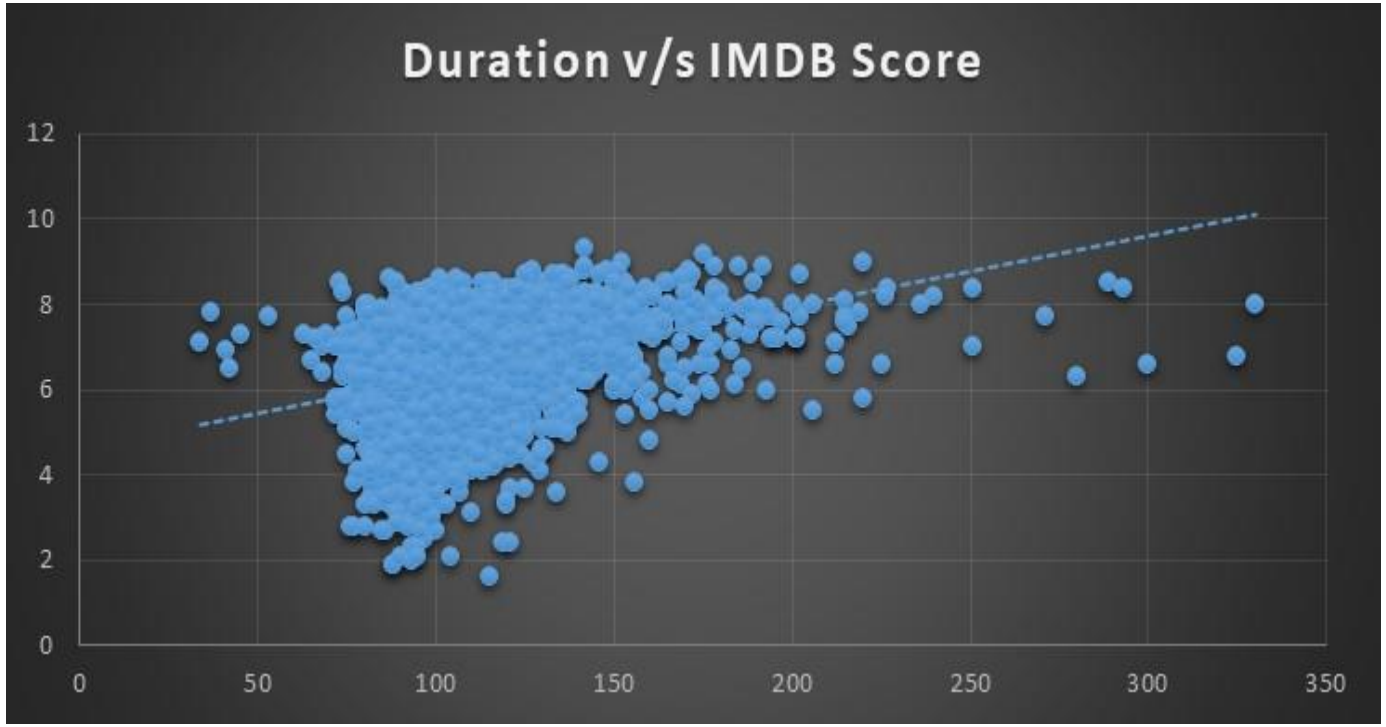
Workings

Calculation of Descriptive Statistics.

duration	
Mean	109.92
Median	106.00
Mode	101.00
Min	34.00
Max	330.00
Range	296.00
Variance	517.61
Standard Deviation	22.75
Correlation Coefficient	0.3604

Workings

Visual representation of Duration and IMDB Score in order to understand the relationship between the same.



Insights

The above scatter plot along with the value of correlation coefficient suggest us that there is a weak positive correlation between movie length and IMDB score.

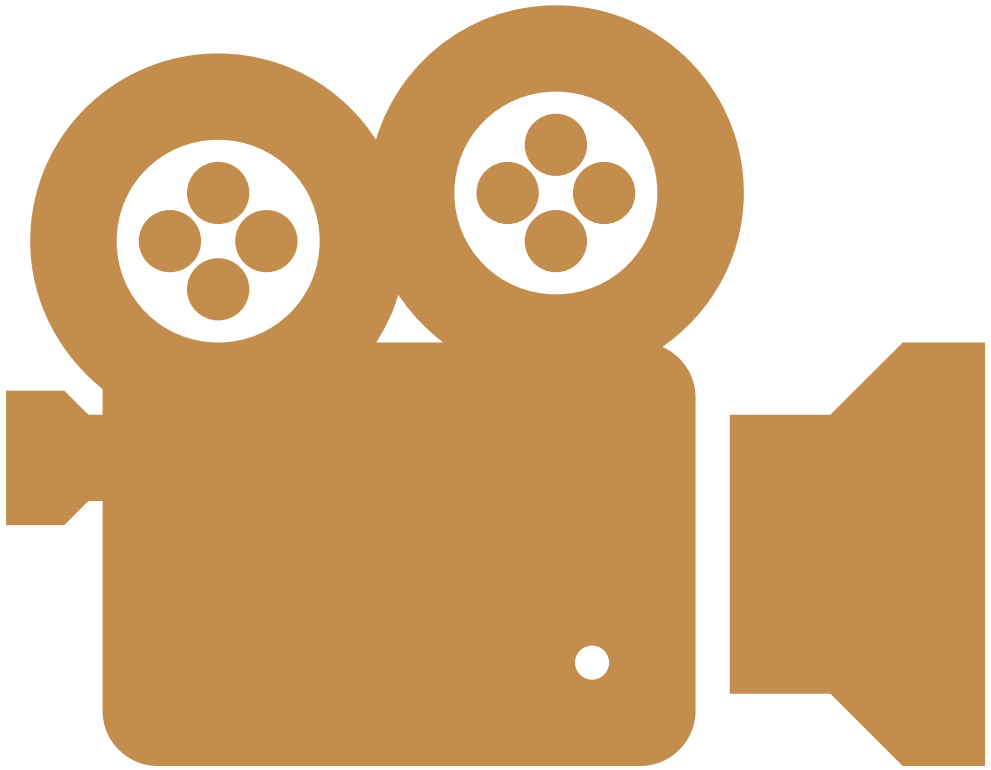
This means that longer movies tend to have slightly higher IMDB scores than shorter movies.

However, there is a lot of variation in the data, and there are many exceptions to this trend also.

Ultimately, it seems that movie length is not a very good predictor of how good a movie will be.

Movie Language Analysis

Task: To determine the most common languages used in movies and analyse their impact on the IMDB score using descriptive statistics.



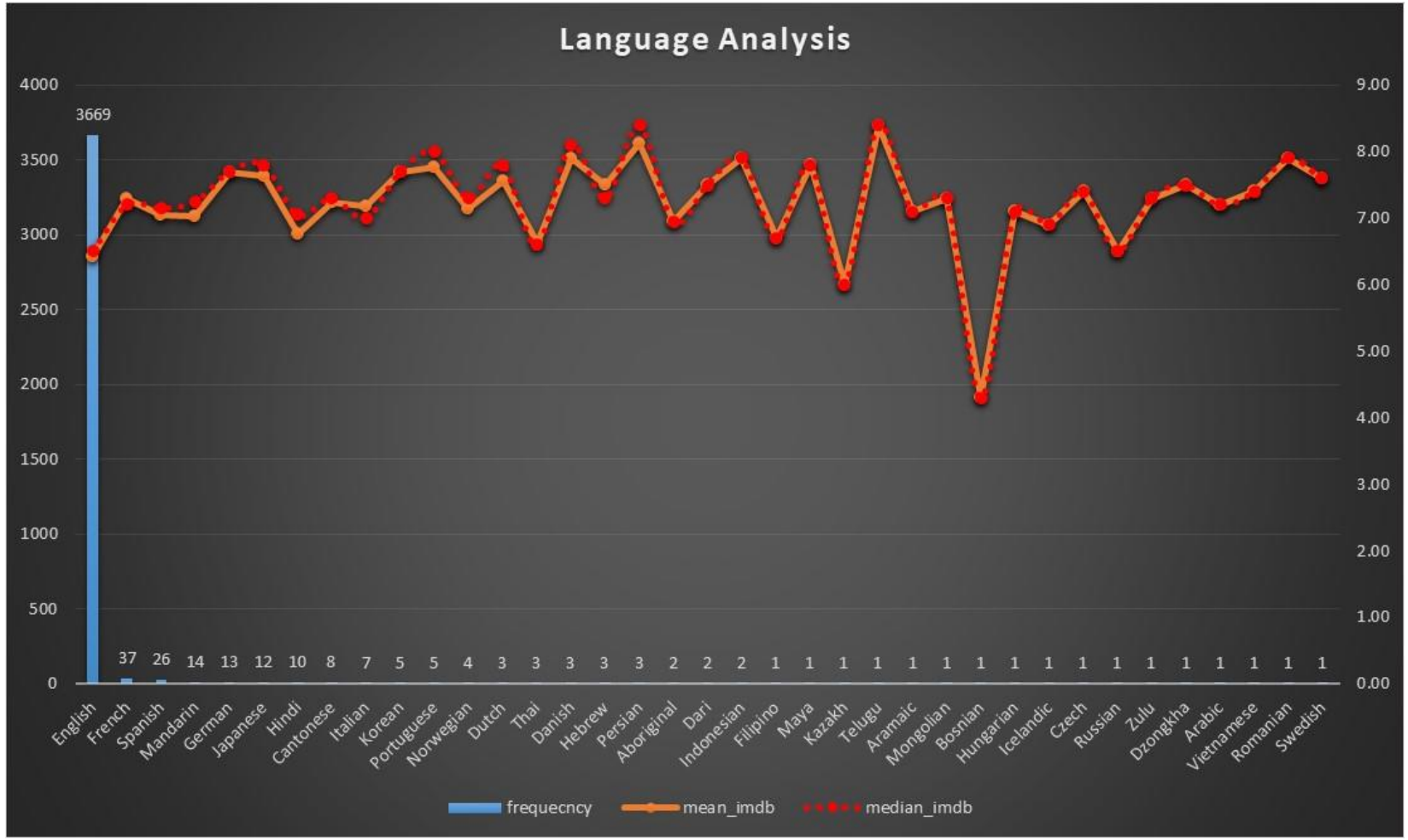
Workings

Calculation of the mean, median, and standard deviation of the IMDB scores for each language.

Language	frequecncy	mean_imdb	median_imdb	stdev_imdb
English	3669	6.42	6.5	1.05
French	37	7.29	7.2	0.56
Spanish	26	7.05	7.15	0.83
Mandarin	14	7.02	7.25	0.77
German	13	7.69	7.7	0.64
Japanese	12	7.63	7.8	0.90
Hindi	10	6.76	7.05	1.11
Cantonese	8	7.24	7.3	0.44
Italian	7	7.19	7	1.16
Korean	5	7.70	7.7	0.57
Portuguese	5	7.76	8	0.98
Norwegian	4	7.15	7.3	0.57
Dutch	3	7.57	7.8	0.40
Thai	3	6.63	6.6	0.45
Danish	3	7.90	8.1	0.53
Hebrew	3	7.50	7.3	0.44
Persian	3	8.13	8.4	0.55
Aboriginal	2	6.95	6.95	0.78
Dari	2	7.50	7.5	0.14
Indonesian	2	7.90	7.9	0.42
Filipino	1	6.70	6.7	NA
Maya	1	7.80	7.8	NA
Kazakh	1	6.00	6	NA
Telugu	1	8.40	8.4	NA
Aramaic	1	7.10	7.1	NA
Mongolian	1	7.30	7.3	NA
Bosnian	1	4.30	4.3	NA
Hungarian	1	7.10	7.1	NA
Icelandic	1	6.90	6.9	NA
Czech	1	7.40	7.4	NA
Russian	1	6.50	6.5	NA
Zulu	1	7.30	7.3	NA
Dzongkha	1	7.50	7.5	NA
Arabic	1	7.20	7.2	NA
Vietnamese	1	7.40	7.4	NA
Romanian	1	7.90	7.9	NA
Swedish	1	7.60	7.6	NA

Workings

Visual representation of Frequency, mean and median.



Insights

- "English" is the most popular language with total of 3669 movies listed and with the mean IMDB score of 6.42. Here, the mean and median values are almost similar with the standard deviation value of 1.05.
- “Telugu” is the language with highest mean IMDB score of 8.40 with only 1 movie listed.
- "French", "Spanish", "Mandarin", "German", "Japanese" are the languages with almost good number of movies listed and also with average IMDB around 7.34 and also with standard deviation below 1.
- "Filipino", "Maya", "Kazakh", "Telugu", "Aramaic", "Mongolian", "Bosnian", "Hungarian", "Icelandic", "Czech", "Russian", "Zulu", "Dzongkha", "Arabic", "Vietnamese", "Romanian", "Swedish" are the languages with lowest listed movies of 1 and the average IMDB of 7.08.
- The lowest IMDB mean is 4.30 which is of "Bosnian" language with only 1 movie listed.
- The correlation coefficient of the frequency and the mean is -0.192968.
- This means that there is a slight tendency for the two variables to move in opposite directions (i.e., when the frequency increases the mean IMDB score decreases), but the relationship is not very strong.

Director Analysis

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

Workings

Calculation of 90th percentile based on the “mean_imdb” (the calculation of “mean_imdb” score of each director is not included here as it exceeds more than 1000 rows. Please refer the attached link at the end of this file to access the Excel file having detailed calculations) of each directors.

Percentile	
90th Percentile	7.5

Workings

- The value of 90th percentile is 7.5. This means 90% of the total average IMDB scores falls below 7.5 and only 10% of the average IMDB scores are above this value.
- I consider the directors who falls under this 10% and who have more than 1 movie listed as the top directors.

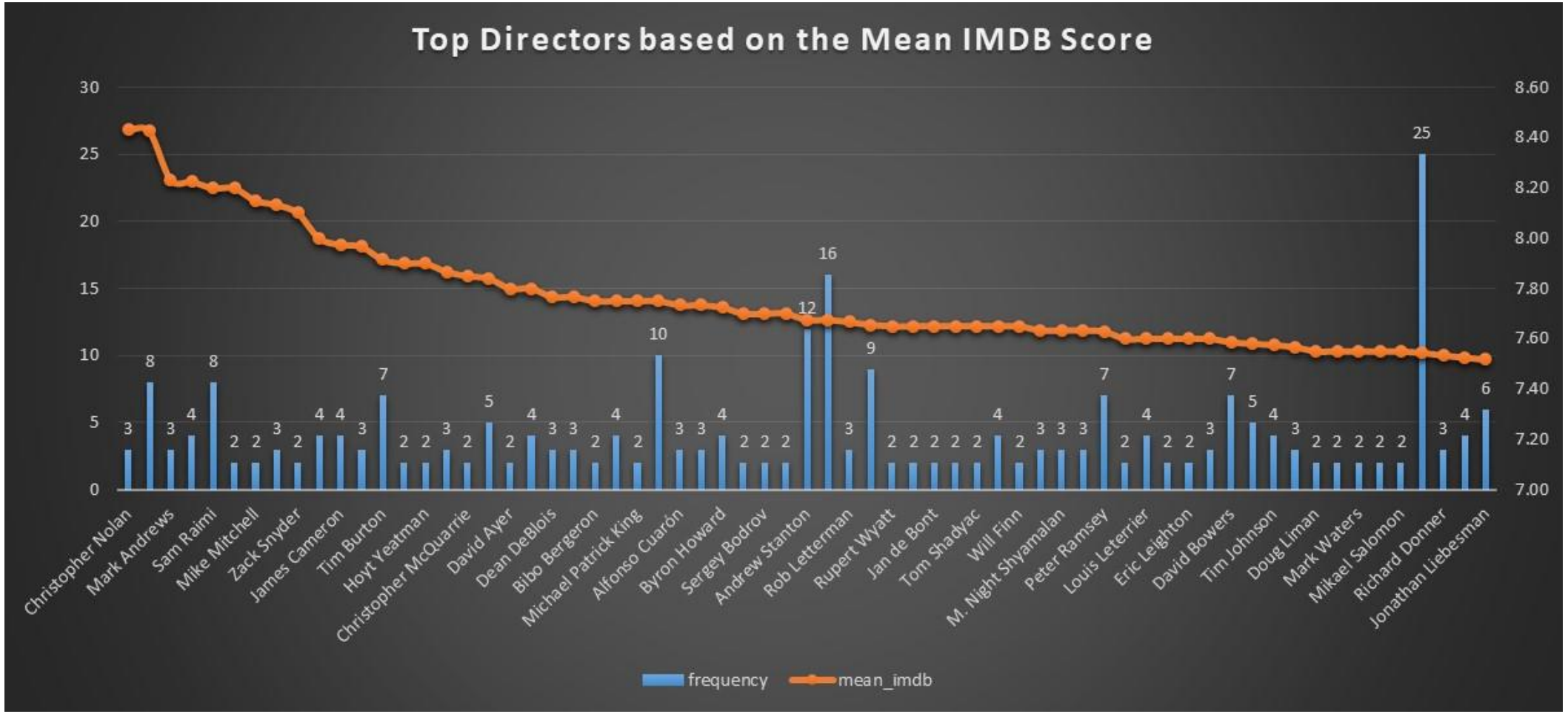
Directors	frequency	mean_imdb
Christopher Nolan	3	8.43
Sam Mendes	8	8.43
Mark Andrews	3	8.23
Peter Jackson	4	8.23
Sam Raimi	8	8.20
Pete Docter	2	8.20
Mike Mitchell	2	8.15
David Yates	3	8.13
Zack Snyder	2	8.10
J.J. Abrams	4	8.00
James Cameron	4	7.98
Don Hall	3	7.97
Tim Burton	7	7.91
Bill Condon	2	7.90
Hoyt Yeatman	2	7.90
Ron Howard	3	7.87
Christopher McQuarrie	2	7.85
Duncan Jones	5	7.84
David Ayer	2	7.80
Michel Gondry	4	7.80
Dean DeBlois	3	7.77
Baz Luhrmann	3	7.77

Directors	frequency	mean_imdb
Bibo Bergeron	2	7.75
Breck Eisner	4	7.75
Michael Patrick King	2	7.75
Nathan Greno	10	7.75
Alfonso Cuarón	3	7.73
McG	3	7.73
Byron Howard	4	7.73
Mimi Leder	2	7.70
Sergey Bodrov	2	7.70
Steve Hickner	2	7.70
Andrew Stanton	12	7.68
Marc Webb	16	7.68
Rob Letterman	3	7.67
Gore Verbinski	9	7.66
Rupert Wyatt	2	7.65
Len Wiseman	2	7.65
Jan de Bont	2	7.65
Kelly Asbury	2	7.65
Tom Shadyac	2	7.65
Wally Pfister	4	7.65
Will Finn	2	7.65
Mike Newell	3	7.63

Directors	frequency	mean_imdb
M. Night Shyamalan	3	7.63
David Soren	3	7.63
Peter Ramsey	7	7.63
James Algar	2	7.60
Louis Leterrier	4	7.60
Vincent Ward	2	7.60
Eric Leighton	2	7.60
Barry Sonnenfeld	3	7.60
David Bowers	7	7.59
Alex Proyas	5	7.58
Tim Johnson	4	7.58
Peter Chelsom	3	7.57
Doug Liman	2	7.55
Bobby Farrelly	2	7.55
Mark Waters	2	7.55
John Moore	2	7.55
Mikael Salomon	2	7.55
Rob Marshall	25	7.54
Richard Donner	3	7.53
Tony Scott	4	7.53
Jonathan Liebesman	6	7.52

Workings

Visual representation of Top Directors.

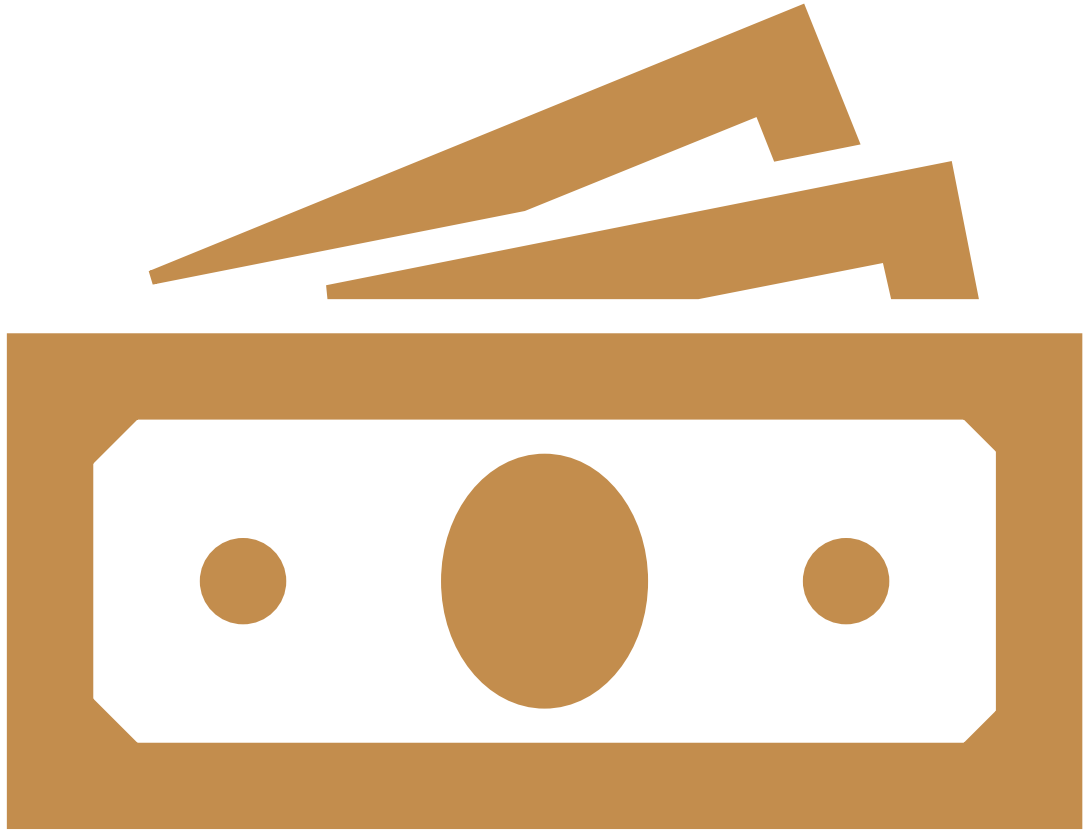


Insights

- "Christopher Nolan", "Sam Mendes" has the highest “mean_imdb” score of 8.43 with 3 and 8 movies listed respectively.
- The maximum number movies listed within the top directors list is 25, which is of director "Rob Marshall" with “mean_imdb” score of 7.54.
- The correlation coefficient value between the frequency and “mean_imdb” is -0.0478, which indicates a slight negative correlation between the same.

Budget Analysis

Task: Understanding the relationship between movie budgets and their financial success.



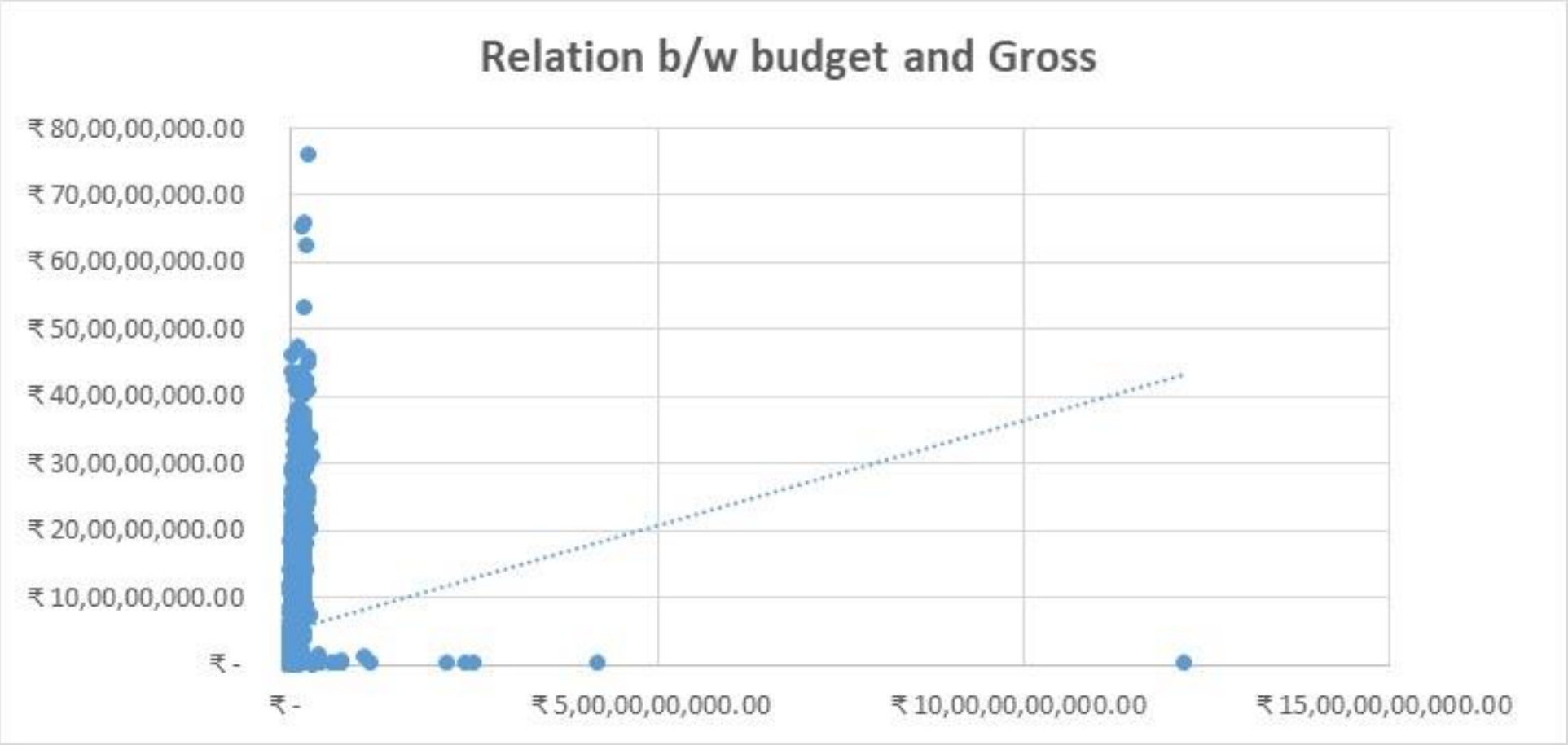
Workings

Calculation of Correlation Coefficient between “Budget” and “Gross Earnings”.

Budget Analysis	
Correlation b/w Budget and Gross Earnings	0.1009

Workings

Scatterplot showing the relationship between Budget and Gross Earnings.



Insights

The correlation coefficient between the budget and the gross earnings of movies is 0.1009.

As the value is almost equals to 0, it is clear that there is no direct relationship between the budget and the gross collection.

Profit Analysis

Task: To calculate the profit margin (gross earnings - budget) for each movie and identify the movies with the highest profit margin.

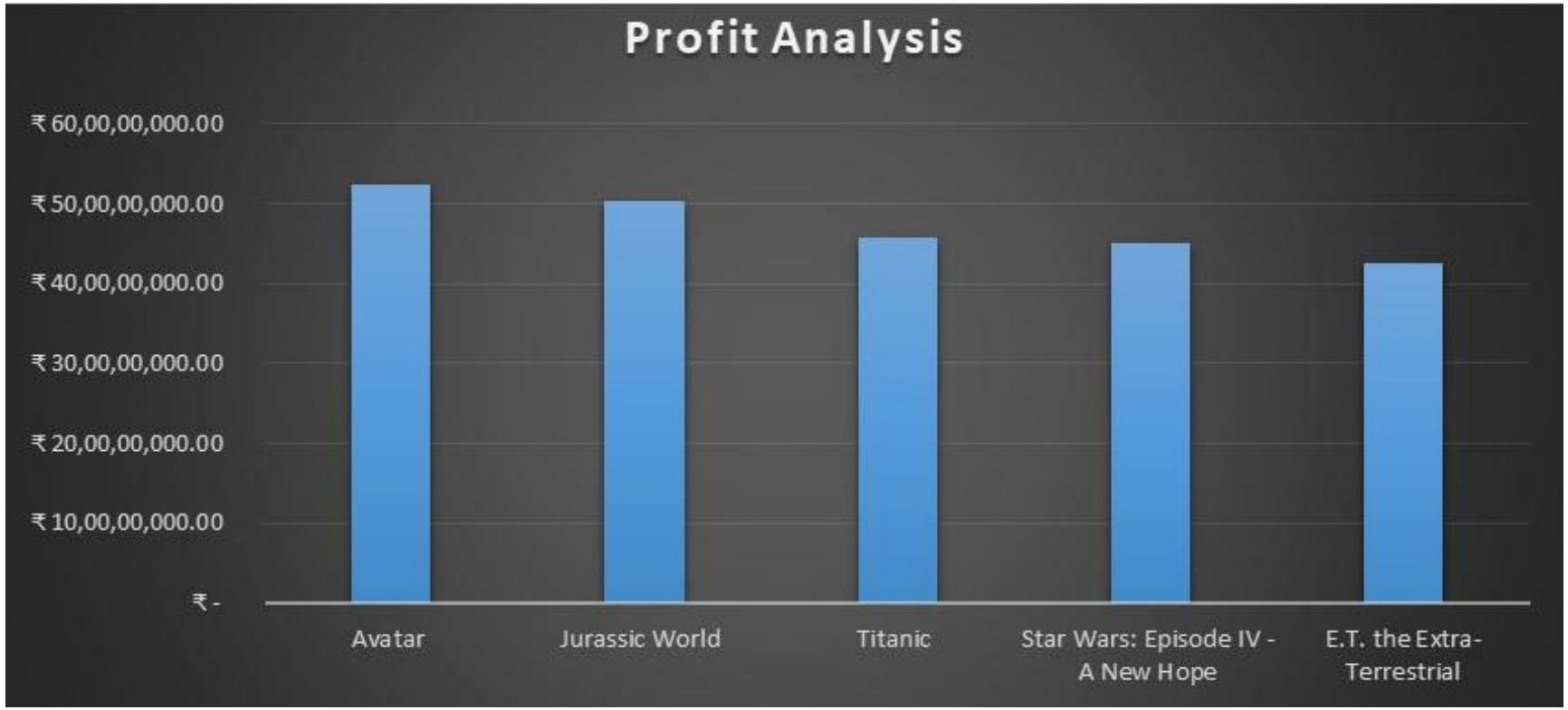
Workings

Calculation of Top 5 movies having highest profits.

Movie Name	Profit Amount	
Avatar	₹	52,35,05,847.00
Jurassic World	₹	50,21,77,271.00
Titanic	₹	45,86,72,302.00
Star Wars: Episode IV - A New Hope	₹	44,99,35,665.00
E.T. the Extra-Terrestrial	₹	42,44,49,459.00

Workings

Visual representation of Top 5 movies having highest profits.



Insights

"Avatar" is the most profitable movie with total profit of Rs.52,35,05,847.

"Jurassic World ","Titanic ","Star Wars: Episode IV - A New Hope ","E.T. the Extra-Terrestrial " are other movies which falls under the top 5 profitable movies.

Critic Analysis

Task: To understand the relationship between “num_critic_for_reviews” and “imdb_score”.

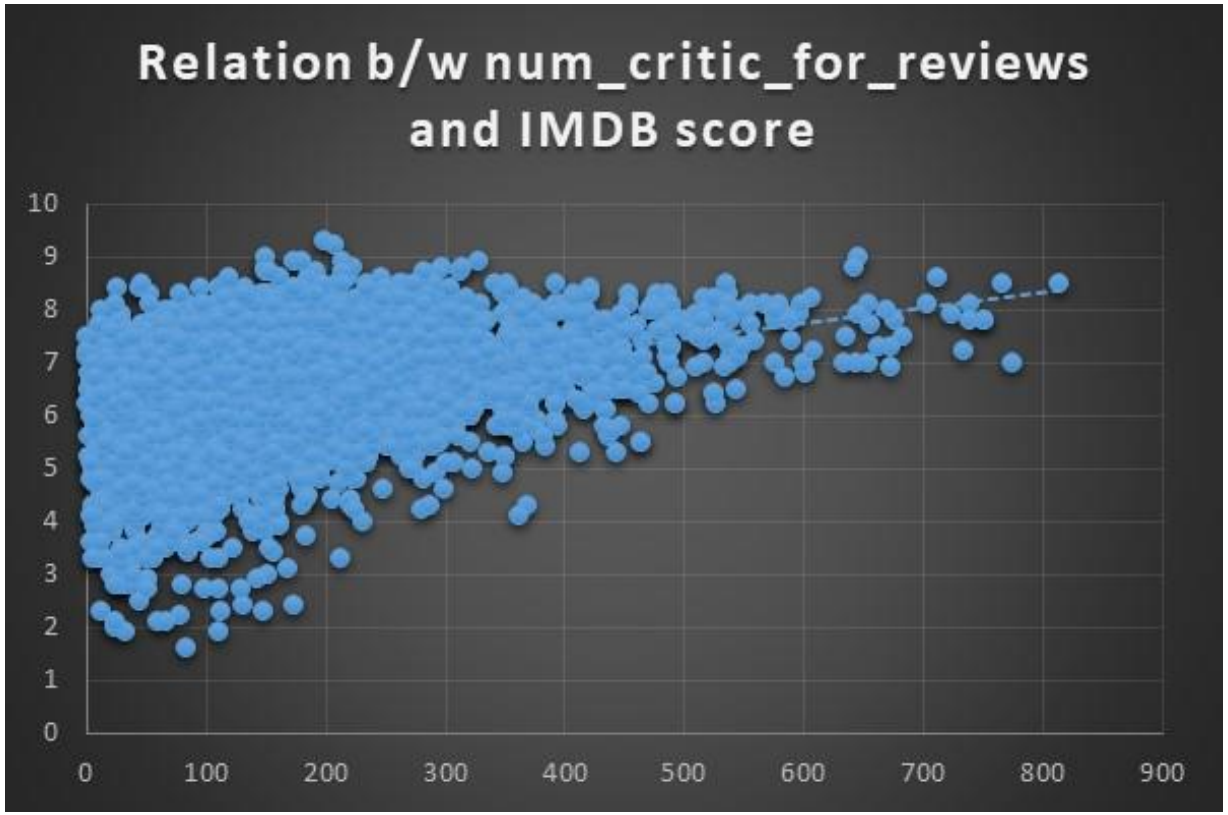
Workings

Calculation of correlation coefficient between “num_critic_for_reviews” and “imdb_score”.

Correlation Coefficient	0.3204

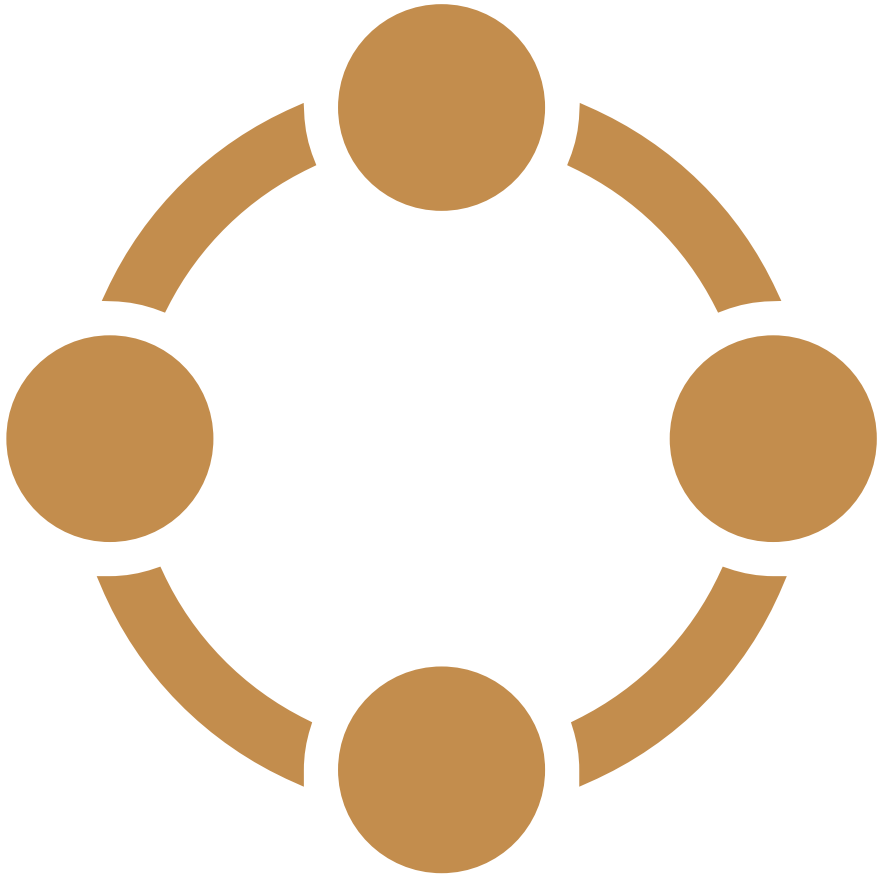
Workings

Scatter plot to show the graphical representation of the relationship between the “num_critic_for_reviews” and “imdb_score”.



Insights

- The correlation coefficient between the “num_critic_for_reviews” and “IMDB score” is 0.3204.
- This, along with the scatter plot, reveals that there is a slight positive correlation between the “critic_for_reviews” and the “imdb score”.



Voted Users Analysis

Task: To identify the relationship between the “number_of_voted_users” and “imdb_score”.

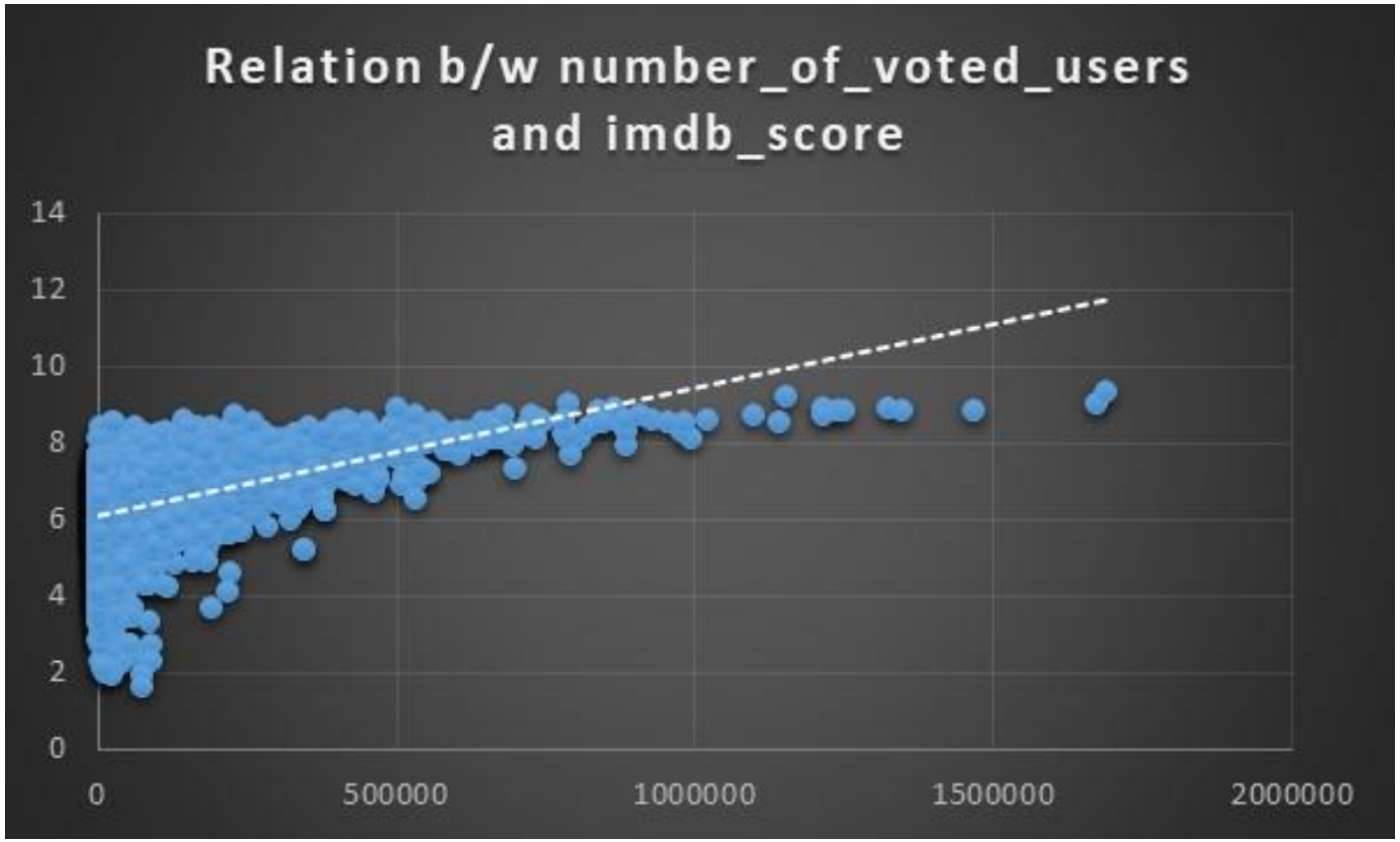
Workings

Calculation of Correlation Coefficient between “number_of_voted_users” and “imdb_score”

Correlation Coefficient	0.4739

Workings

Scatter plot visualizing the relationship between the “number_of_voted_users” and “imdb_score”.



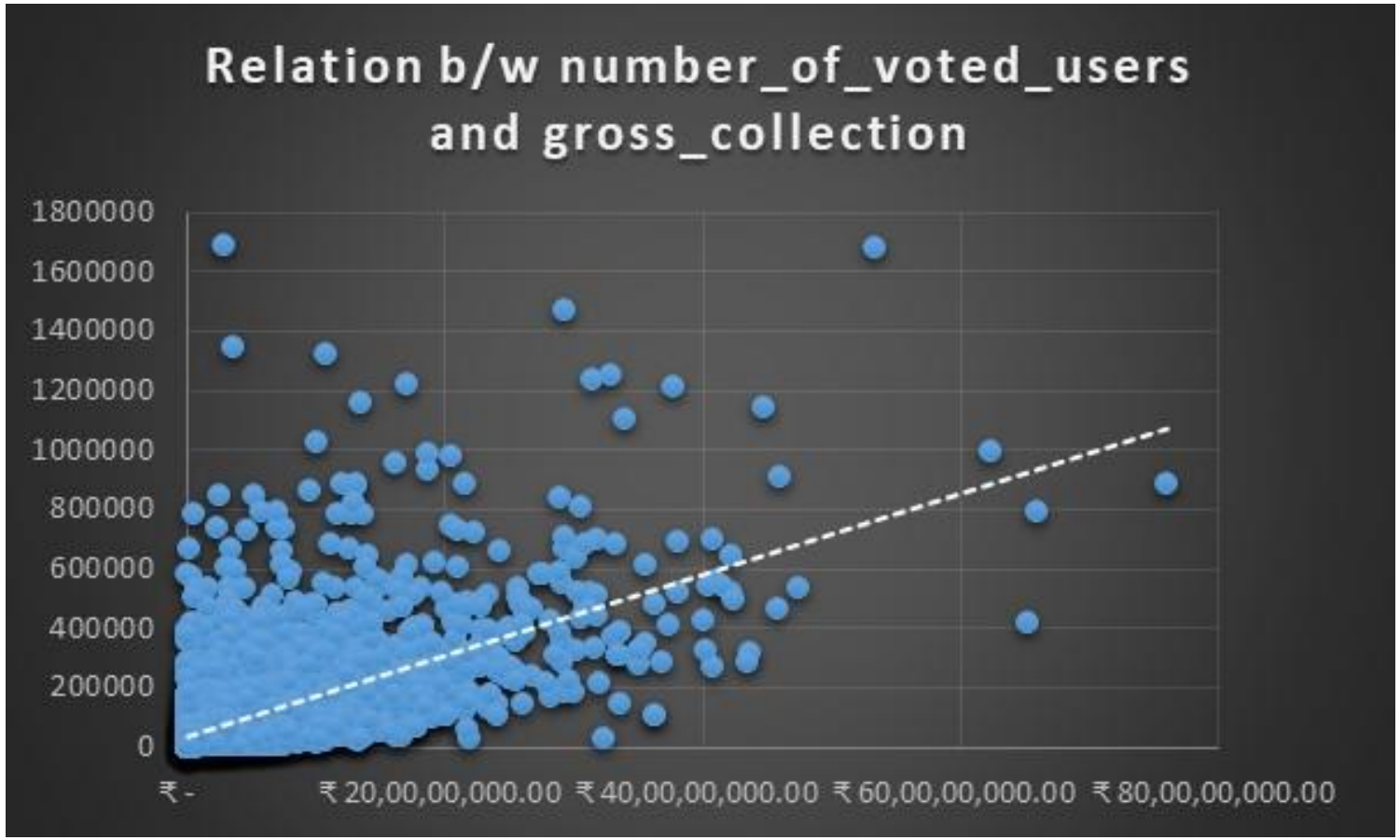
Workings

Calculation of Correlation Coefficient between “number_of_voted_users” and “gross_collection”.

Correlation Coefficient	0.6290

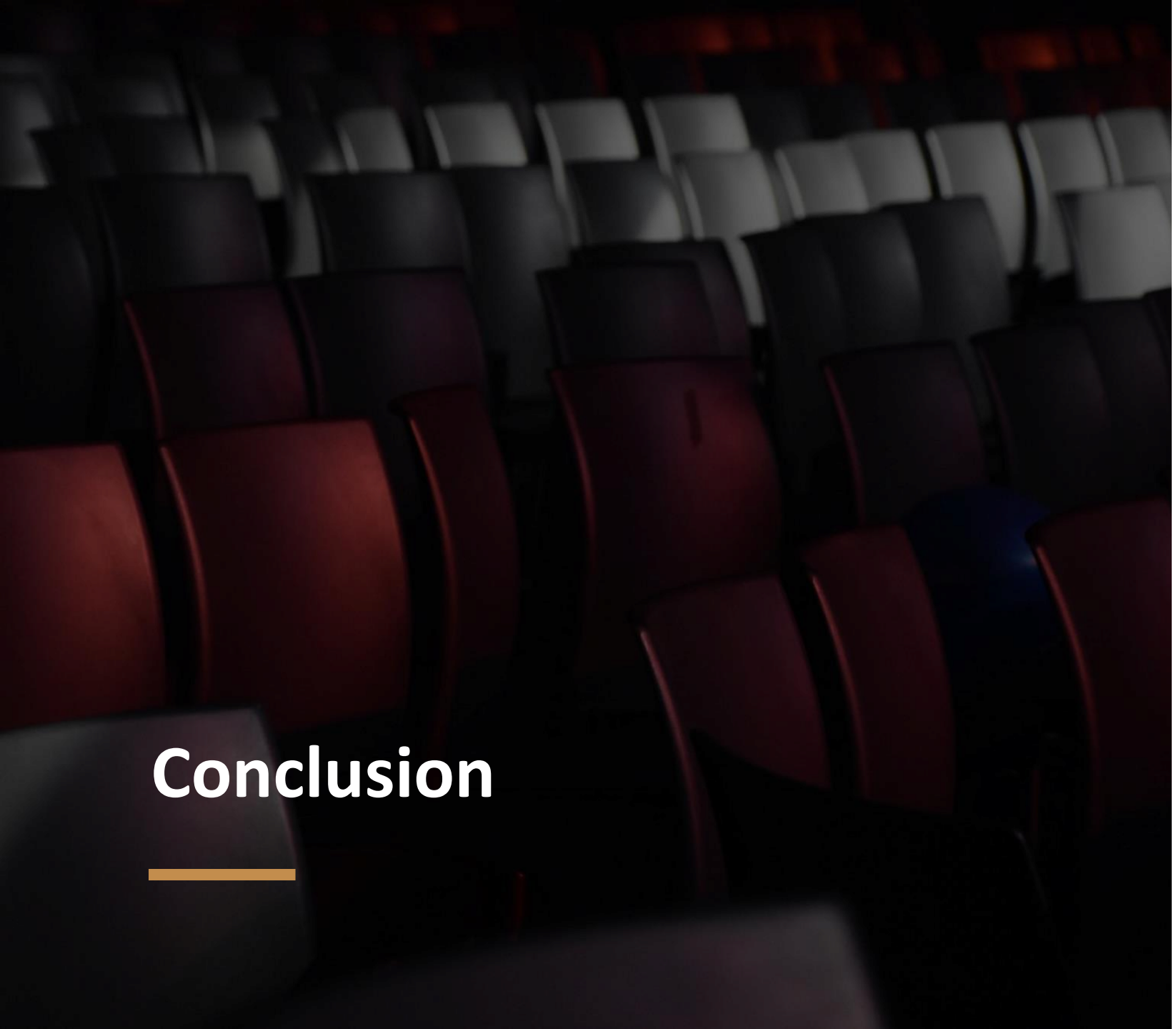
Workings

Scatter plot visualizing the “number_of_voted_users” and “gross_collection”.



Insights

- Based on the Correlation Coefficient and the scatter plot, it is evident that there is a strong positive correlation between the “number_of_voted_users” and “imdb_score”.
- Also, the relationship between the “nubmer_of_voted_users” and the “gross_collection” is direct and strongly positive.



Conclusion

Based on the above analysis, it is clear that the success of a movie is dependent on, not just one but, many factors such as budget, genre, duration, language, director, actors, number of people voted etc.

THANK YOU

Feel free to access the excel file that I worked : Link

sathishgudri@gmail.com