Trainity Data Analytics Training

Project 5

IMDB Movie Analysis

Date: 17/12/24 **Name-Sudipta Samanta**

Project Description:

This project involves an analysis of a dataset related to IMDB movies, with the goal of understanding the factors that influence a movie's success, as measured by its IMDB score. By exploring various attributes such as movie genre, duration, language, director, and budget, the project aims to uncover patterns and relationships that provide insights into what makes a movie successful. The analysis will be done using Microsoft Excel, with visualizations and descriptive statistics to identify trends and provide actionable insights for movie producers, directors, and investors.

Approach:

1. Data Cleaning: The initial step will involve preprocessing the dataset to handle missing values, duplicates, and any inconsistencies. The data types will be adjusted where necessary, and new features may be engineered for more granular insights. This will ensure the dataset is ready for analysis.

2. Data Analysis: Each task will focus on a specific aspect of the dataset, as outlined in the problem statement. Descriptive statistics will be calculated for each factor (such as movie genre, duration, and budget), and visualizations will be created to explore relationships between these factors and IMDB scores. The analysis will be performed using Excel's functions like AVERAGE, MEDIAN, STDEV, COUNTIF, CORREL, and others.

3. Five 'Whys' Approach: For deeper insights, the "Five Whys" technique will be used to explore root causes behind observed patterns. This will allow for a more in-depth understanding of how and why certain factors impact a movie's success.

4. Visualization: Charts and graphs will be used to present findings, such as scatter plots for movie duration and IMDB score, pie charts for genre distribution, and bar graphs for language distribution. A trendline will be added to scatter plots to assess the strength of relationships.

5. Report and Data Story: A comprehensive report will be prepared to present the findings of the analysis. This report will tell a story, including an overview of the dataset, the steps followed, key insights derived from the data, and recommendations for movie producers and investors.

Tech-Stack Used:

Microsoft Excel 365: Used for data cleaning, analysis, and visualization. Key features used include functions like AVERAGE, MEDIAN, STDEV, VAR, CORREL, and PERCENTILE. Pivot tables and charts will also be used for summarization and visualization.

Insights:

1. Movie Genre and IMDB Score: Different genres will likely have varying average IMDB scores. By analyzing the genre distribution and calculating descriptive statistics for each genre, insights can be gained into which genres tend to perform better in terms of ratings. For example, genres like drama or action may consistently score higher, while others like horror or fantasy may have more varied ratings.

2. Movie Duration and IMDB Score: The relationship between movie duration and IMDB score will be explored. Longer movies may receive higher ratings due to their ability to develop complex plots, but there could also be a threshold after which longer durations lead to lower ratings due to viewer fatigue. This insight could be valuable for filmmakers considering the optimal length for their films.

3. Language and Movie Ratings: An analysis of movies by language will reveal whether movies in certain languages (e.g., English, Spanish, or Hindi) tend to have higher or lower IMDB ratings. This might reflect global audience preferences and the international appeal of movies in specific languages.

4. Director Influence on Success: Directors with a strong track record of high IMDB ratings will be identified. The correlation between a director’s average IMDB score and the success of their movies could guide producers in their choice of directors. This may also reveal patterns like certain directors consistently delivering successful movies across genres.

5. Budget and Profit Margin: The relationship between movie budgets and gross earnings will be analyzed to identify movies with the highest profit margins. This analysis will provide insight into how budget influences not only the quality of a movie but also its financial success, offering actionable recommendations for investment in future projects.

Through these insights, the project will help stakeholders understand the key drivers behind movie success, allowing them to make data-driven decisions for future productions.

<u>RESULTS</u> :-

A. **Movie Genre Analysis:** Analyze the distribution of movie genres and their impact on the IMDB score.

- **Task:** 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.

Query:

For no of movies: =COUNTIF(D$2:D$3787,"*" & N7 & "*")

For mean: =AVERAGEIF(D$2:D$3787,"*" & N7 & "*",I$2:I$3787)

For median: =MEDIAN(IF(ISNUMBER(SEARCH("*" & N7 & "*",D$2:D$3787)),I$2:I$3787))

For mode: =MODE(IF(ISNUMBER(SEARCH("*" & N7 & "*",D$2:D$3787)),I$2:I$3787))

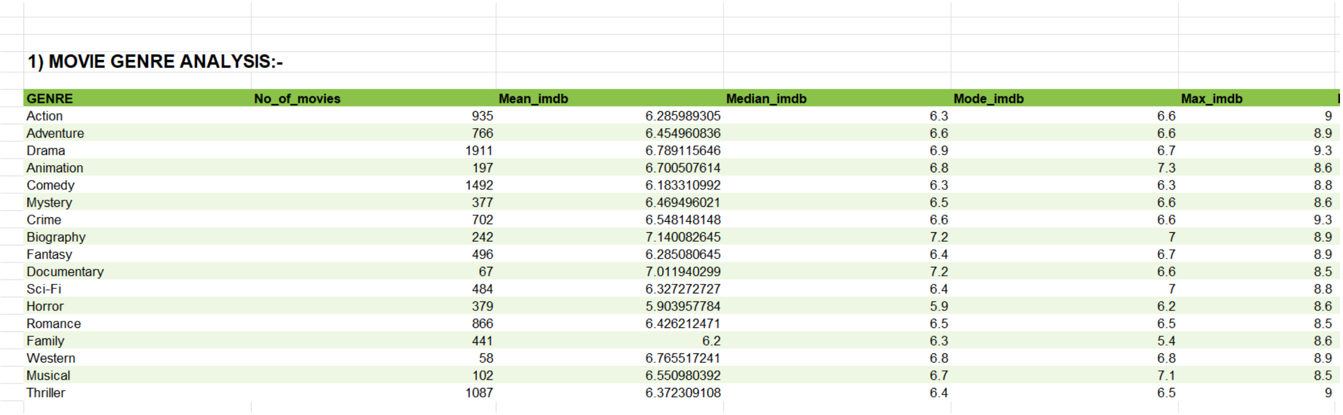
Max IMDB: =MAXIFS(I$2:I$3787,D$2:D$3787,"*" & N7 & "*")

Min IMDB: =MINIFS(I$2:I$3787,D$2:D$3787,"*" & N7 & "*")

Standard deviation IMDB: =STDEV(IF(ISNUMBER(SEARCH("*" & N7 & "*",D$2:D$3787)),I$2:I$3787))

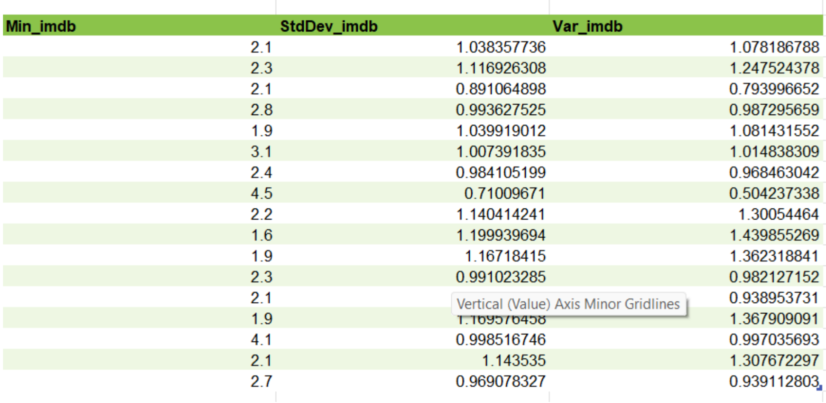
Variation IMDB: =VAR(IF(ISNUMBER(SEARCH("*" & N7 & "*",D$2:D$3787)),I$2:I$3787))

Output:



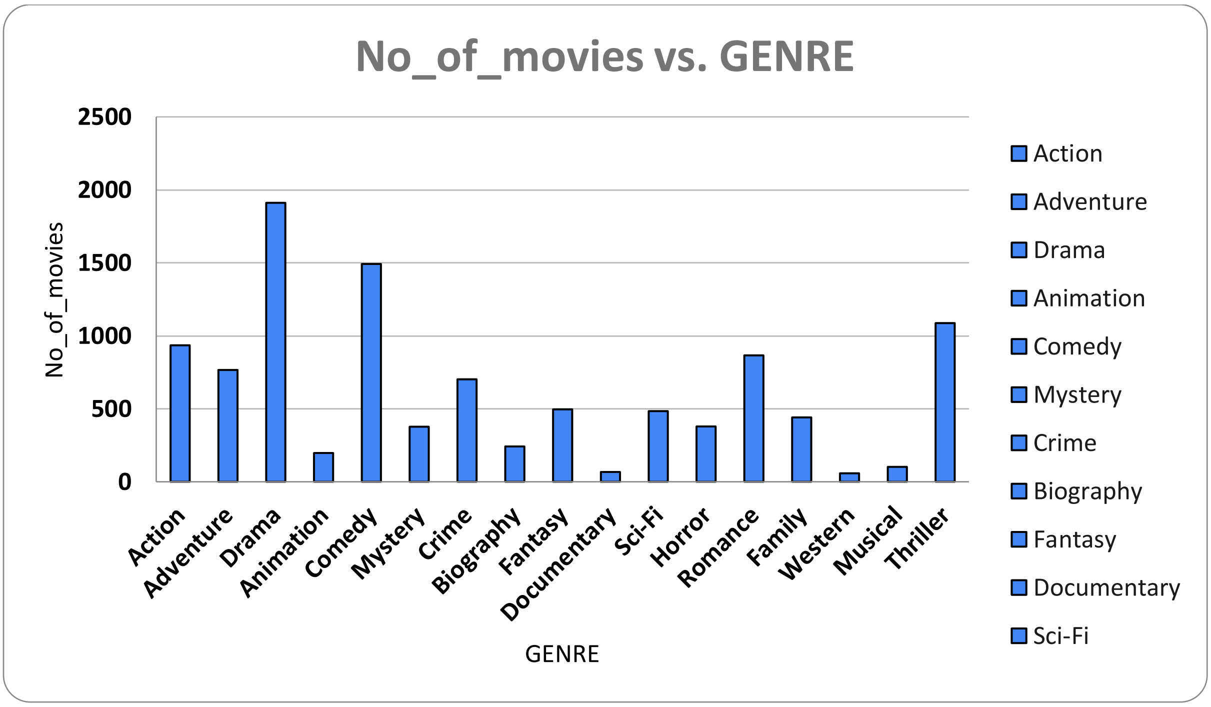
1) MOVIE GENRE ANALYSIS:-

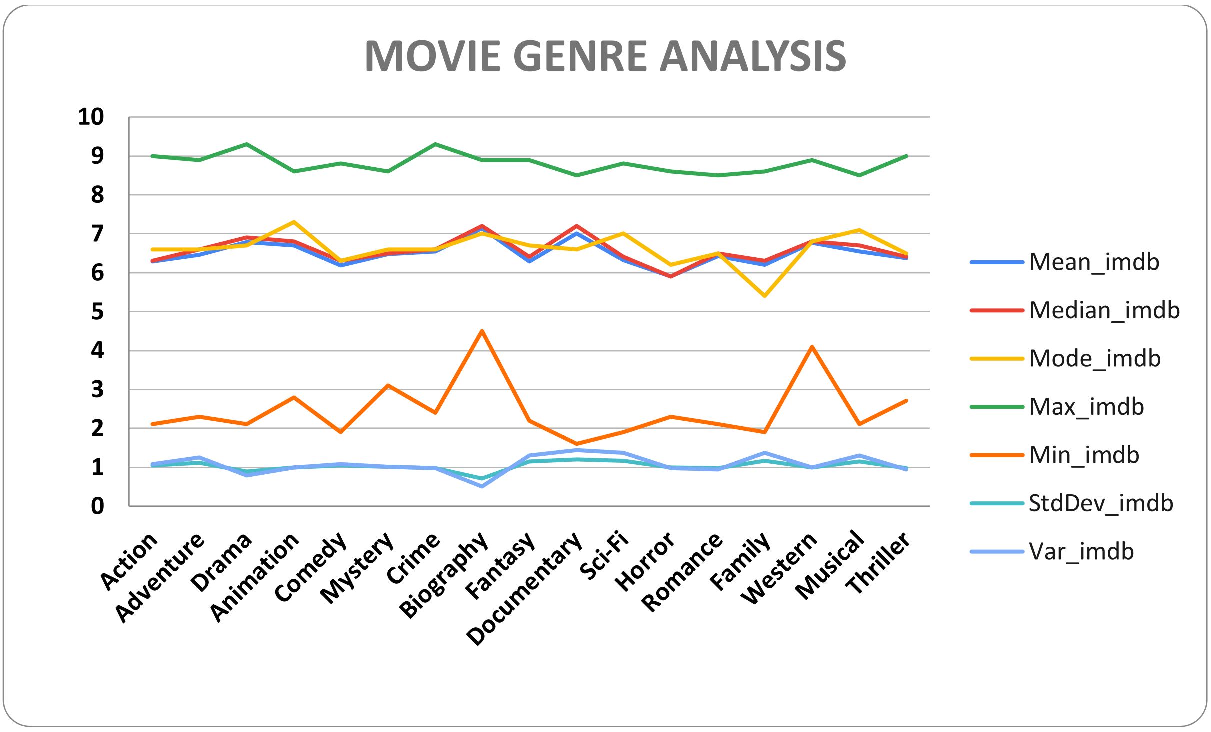
GENRE	No_of_movies	Mean_imdb	Median_imdb	Mode_imdb	Max_imdb
Action	935	6.285989305	6.3	6.6	9
Adventure	766	6.454960836	6.6	6.6	8.9
Drama	1911	6.789115646	6.9	6.7	9.3
Animation	197	6.700507614	6.8	7.3	8.6
Comedy	1492	6.183310992	6.3	6.3	8.8
Mystery	377	6.469496021	6.5	6.6	8.6
Crime	702	6.548148148	6.6	6.6	9.3
Biography	242	7.140082645	7.2	7	8.9
Fantasy	496	6.285080645	6.4	6.7	8.9
Documentary	67	7.011940299	7.2	6.6	8.5
Sci-Fi	484	6.327272727	6.4	7	8.8
Horror	379	5.903957784	5.9	6.2	8.6
Romance	866	6.426212471	6.5	6.5	8.5
Family	441	6.2	6.3	5.4	8.6
Western	58	6.765517241	6.8	6.8	8.9
Musical	102	6.550980392	6.7	7.1	8.5
Thriller	1087	6.372309108	6.4	6.5	9



Min_imdb	StdDev_imdb	Var_imdb
2.1	1.038357736	1.078186788
2.3	1.116926308	1.247524378
2.1	0.891064898	0.793996652
2.8	0.993627525	0.987295659
1.9	1.039919012	1.081431552
3.1	1.007391835	1.014838309
2.4	0.984105199	0.968463042
4.5	0.71009671	0.504237338
2.2	1.140414241	1.30054464
1.6	1.199939694	1.439855269
1.9	1.16718415	1.362318841
2.3	0.991023285	0.982127152
2.1	[illegible]	0.938953731
1.9	1.169576458	1.367909091
4.1	0.998516746	0.997035693
2.1	1.143535	1.307672297
2.7	0.969078327	0.939112803

Vertical (Value) Axis Minor Gridlines





Most Common Genre is Drama.

B. Movie Duration Analysis: Analyze the distribution of movie durations and its impact on the IMDB score.

- **Task**: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Query:

For mean: =AVERAGE(B2:B3787)

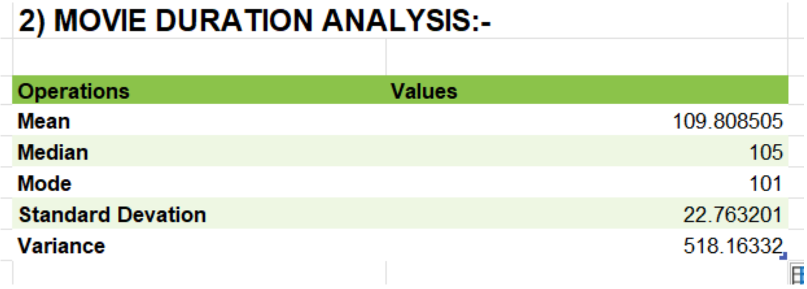
For median: =MEDIAN(B2:B3787)

For mode: =MODE(B2:B3787)

Standard Deviation: =STDEV(B2:B3787)

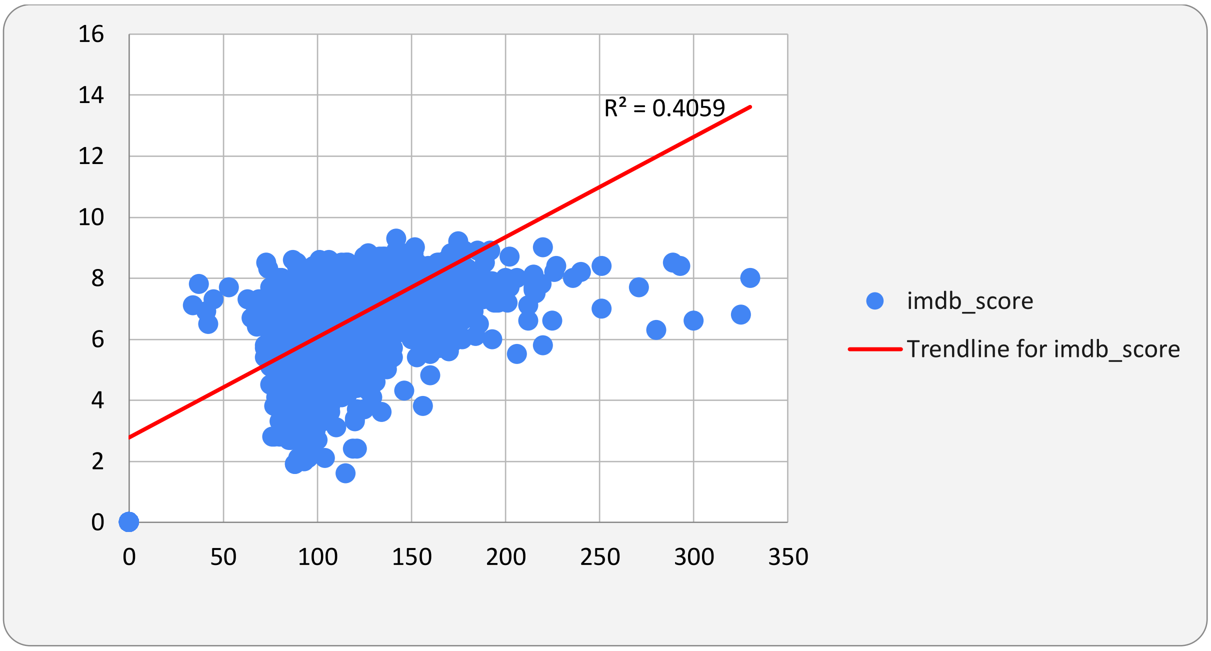
Variation: =VAR(B:B)

Output:



2) MOVIE DURATION ANALYSIS:-

Operations	Values
Mean	109.808505
Median	105
Mode	101
Standard Devation	22.763201
Variance	518.16332



C. Language Analysis: Situation: Examine the distribution of movies based on their language.

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

Query:

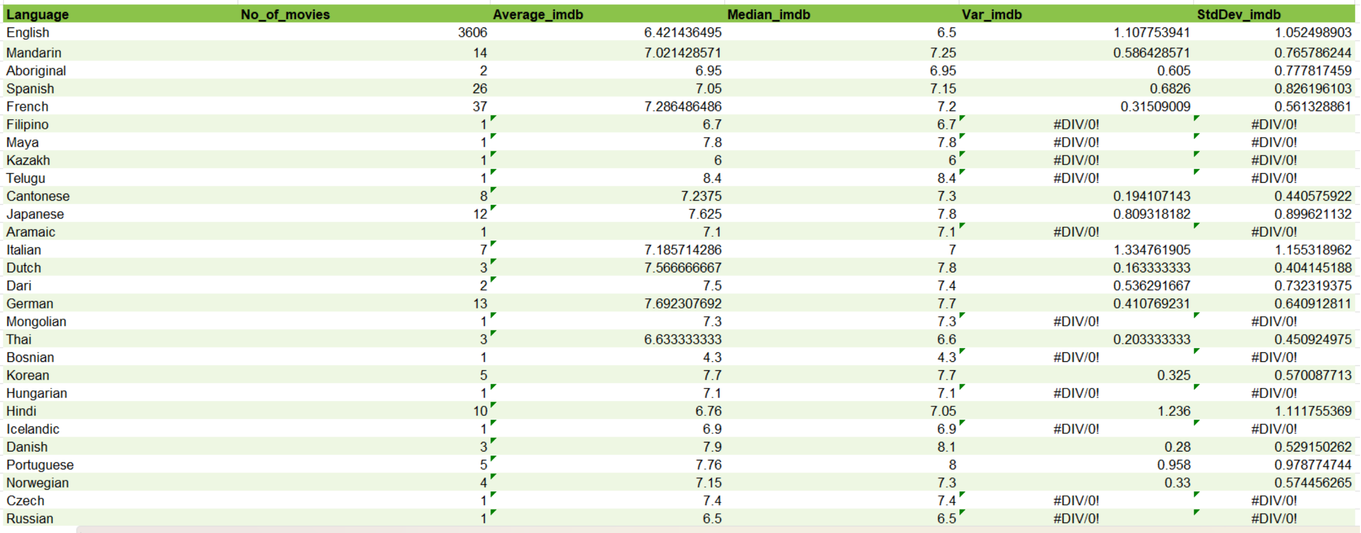
For counting no of movies: =COUNTIF(F$2:F$3787,N104)

For mean: =AVERAGEIF(F2:F3787,N104,I2:I3787)

For median: =MEDIAN(IF(ISNUMBER(SEARCH("*" & N104 & "*",F$2:F$3787)),I$2:I$3787))

Standard Deviation =STDEV(IF(ISNUMBER(SEARCH("*" & M104 & "*",E$2:E$3788)),H$2:H$3788))

Output:



Language	No_of_movies	Average_imdb	Median_imdb	Var_imdb	StdDev_imdb
English	3606	6.421436495	6.5	1.107753941	1.052498903
Mandarin	14	7.021428571	7.25	0.586428571	0.765786244
Aboriginal	2	6.95	6.95	0.605	0.777817459
Spanish	26	7.05	7.15	0.6826	0.826196103
French	37	7.286486486	7.2	0.31509009	0.561328861
Filipino	1	6.7	6.7	#DIV/0!	#DIV/0!
Maya	1	7.8	7.8	#DIV/0!	#DIV/0!
Kazakh	1	6	6	#DIV/0!	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!	#DIV/0!
Cantonese	8	7.2375	7.3	0.194107143	0.440575922
Japanese	12	7.625	7.8	0.809318182	0.899621132
Aramaic	1	7.1	7.1	#DIV/0!	#DIV/0!
Italian	7	7.185714286	7	1.334761905	1.155318962
Dutch	3	7.566666667	7.8	0.163333333	0.404145188
Dari	2	7.5	7.4	0.536291667	0.732319375
German	13	7.692307692	7.7	0.410769231	0.640912811
Mongolian	1	7.3	7.3	#DIV/0!	#DIV/0!
Thai	3	6.633333333	6.6	0.203333333	0.450924975
Bosnian	1	4.3	4.3	#DIV/0!	#DIV/0!
Korean	5	7.7	7.7	0.325	0.570087713
Hungarian	1	7.1	7.1	#DIV/0!	#DIV/0!
Hindi	10	6.76	7.05	1.236	1.111755369
Icelandic	1	6.9	6.9	#DIV/0!	#DIV/0!
Danish	3	7.9	8.1	0.28	0.529150262
Portuguese	5	7.76	8	0.958	0.978774744
Norwegian	4	7.15	7.3	0.33	0.574456265
Czech	1	7.4	7.4	#DIV/0!	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!	#DIV/0!

Most commom Language used in movies : English

D. Director Analysis: Influence of directors on movie ratings.

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

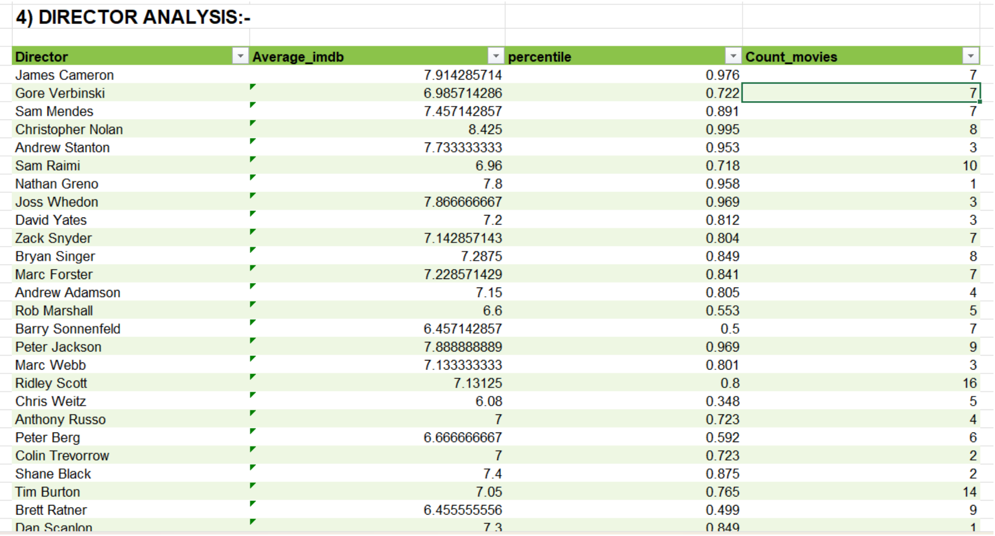
  Query:

  =AVERAGEIF(A2:A3787,N149,I2:I3787)

  =PERCENTRANK.EXC(O$149:O$1899,O149)

  =COUNTIF(A$2:A$3787,N149)

  Output:



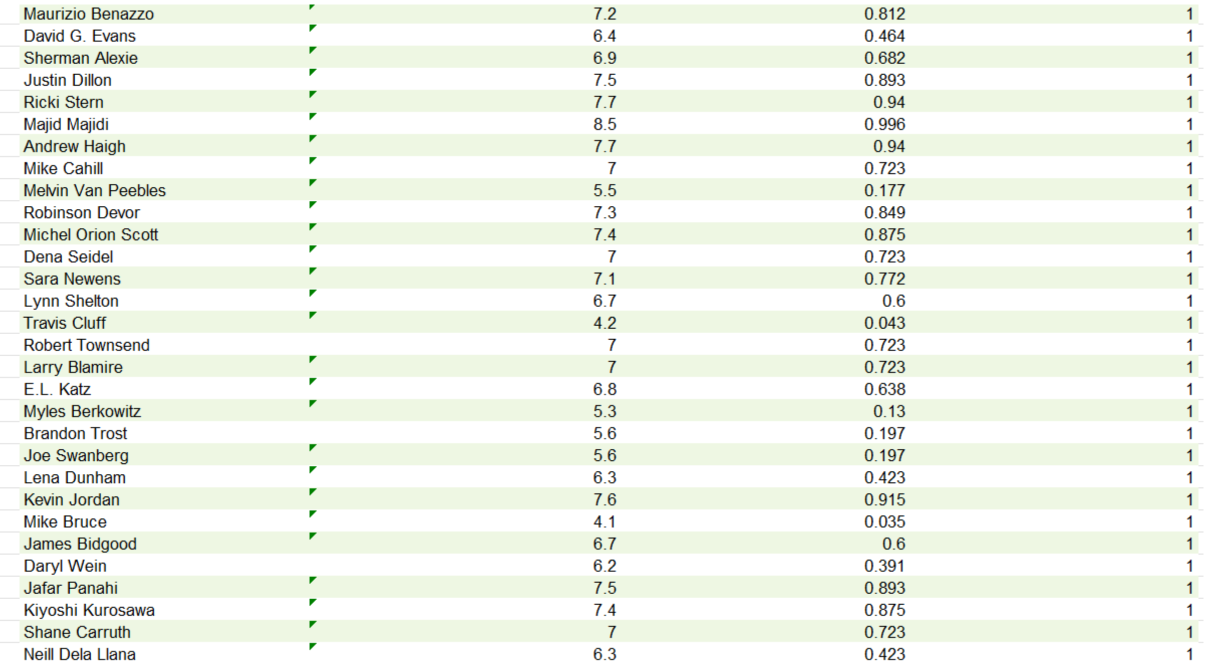
4) DIRECTOR ANALYSIS:-

Director	Average_imdb	percentile	Count_movies
James Cameron	7.914285714	0.976	7
Gore Verbinski	6.985714286	0.722	7
Sam Mendes	7.457142857	0.891	7
Christopher Nolan	8.425	0.995	8
Andrew Stanton	7.733333333	0.953	3
Sam Raimi	6.96	0.718	10
Nathan Greno	7.8	0.958	1
Joss Whedon	7.866666667	0.969	3
David Yates	7.2	0.812	3
Zack Snyder	7.142857143	0.804	7
Bryan Singer	7.2875	0.849	8
Marc Forster	7.228571429	0.841	7
Andrew Adamson	7.15	0.805	4
Rob Marshall	6.6	0.553	5
Barry Sonnenfeld	6.457142857	0.5	7
Peter Jackson	7.888888889	0.969	9
Marc Webb	7.133333333	0.801	3
Ridley Scott	7.13125	0.8	16
Chris Weitz	6.08	0.348	5
Anthony Russo	7	0.723	4
Peter Berg	6.666666667	0.592	6
Colin Trevorrow	7	0.723	2
Shane Black	7.4	0.875	2
Tim Burton	7.05	0.765	14
Brett Ratner	6.455555556	0.499	9
Dan Scanlon	7.3	0.849	1

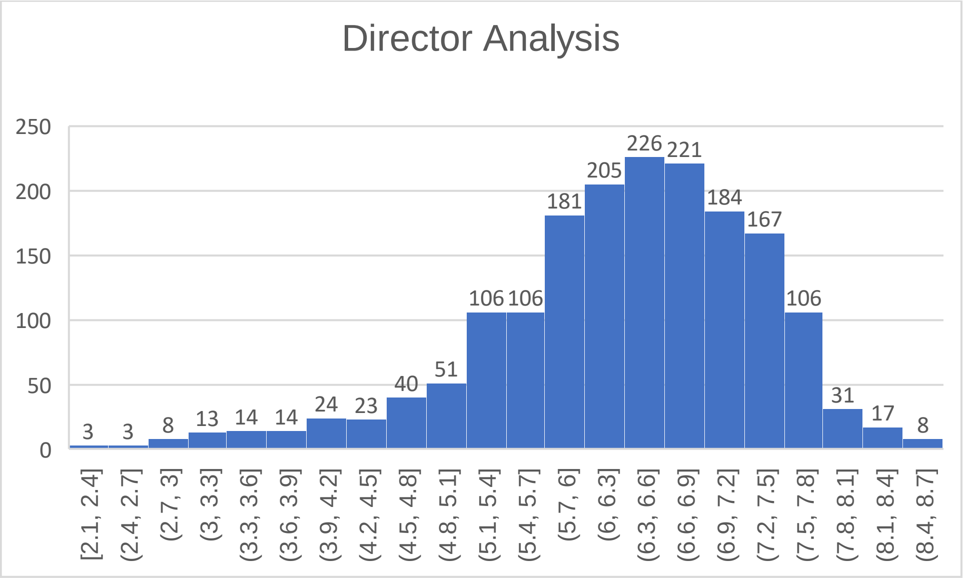
-

-

-



Maurizio Benazzo	7.2	0.812	1
David G. Evans	6.4	0.464	1
Sherman Alexie	6.9	0.682	1
Justin Dillon	7.5	0.893	1
Ricki Stern	7.7	0.94	1
Majid Majidi	8.5	0.996	1
Andrew Haigh	7.7	0.94	1
Mike Cahill	7	0.723	1
Melvin Van Peebles	5.5	0.177	1
Robinson Devor	7.3	0.849	1
Michel Orion Scott	7.4	0.875	1
Dena Seidel	7	0.723	1
Sara Newens	7.1	0.772	1
Lynn Shelton	6.7	0.6	1
Travis Cluff	4.2	0.043	1
Robert Townsend	7	0.723	1
Larry Blamire	7	0.723	1
E.L. Katz	6.8	0.638	1
Myles Berkowitz	5.3	0.13	1
Brandon Trost	5.6	0.197	1
Joe Swanberg	5.6	0.197	1
Lena Dunham	6.3	0.423	1
Kevin Jordan	7.6	0.915	1
Mike Bruce	4.1	0.035	1
James Bidgood	6.7	0.6	1
Daryl Wein	6.2	0.391	1
Jafar Panahi	7.5	0.893	1
Kiyoshi Kurosawa	7.4	0.875	1
Shane Carruth	7	0.723	1
Neill Dela Llana	6.3	0.423	1



E. Budget Analysis: Explore the relationship between movie budgets and their financial success.

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

Query:

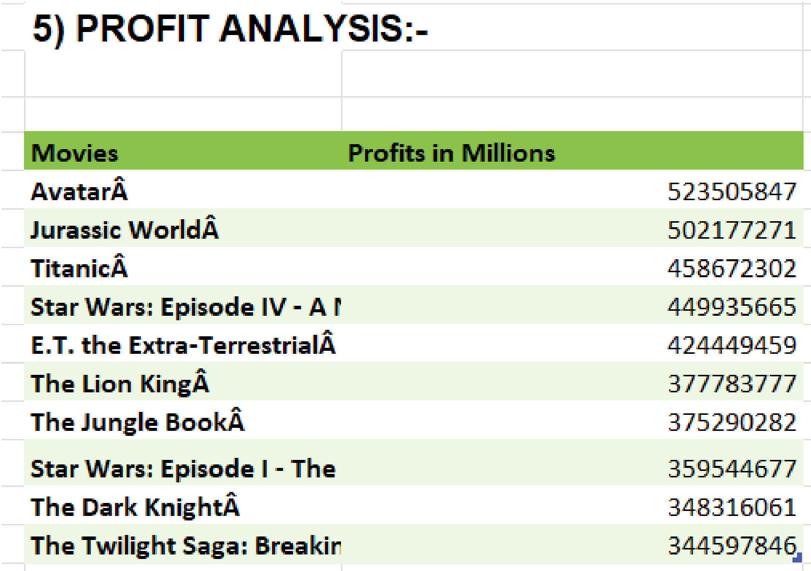
=C2-H2

We can use CORREL function to calculate correlation coefficients between movie budgets and gross earnings.

We can MAX function to get highest profit margin and we can use

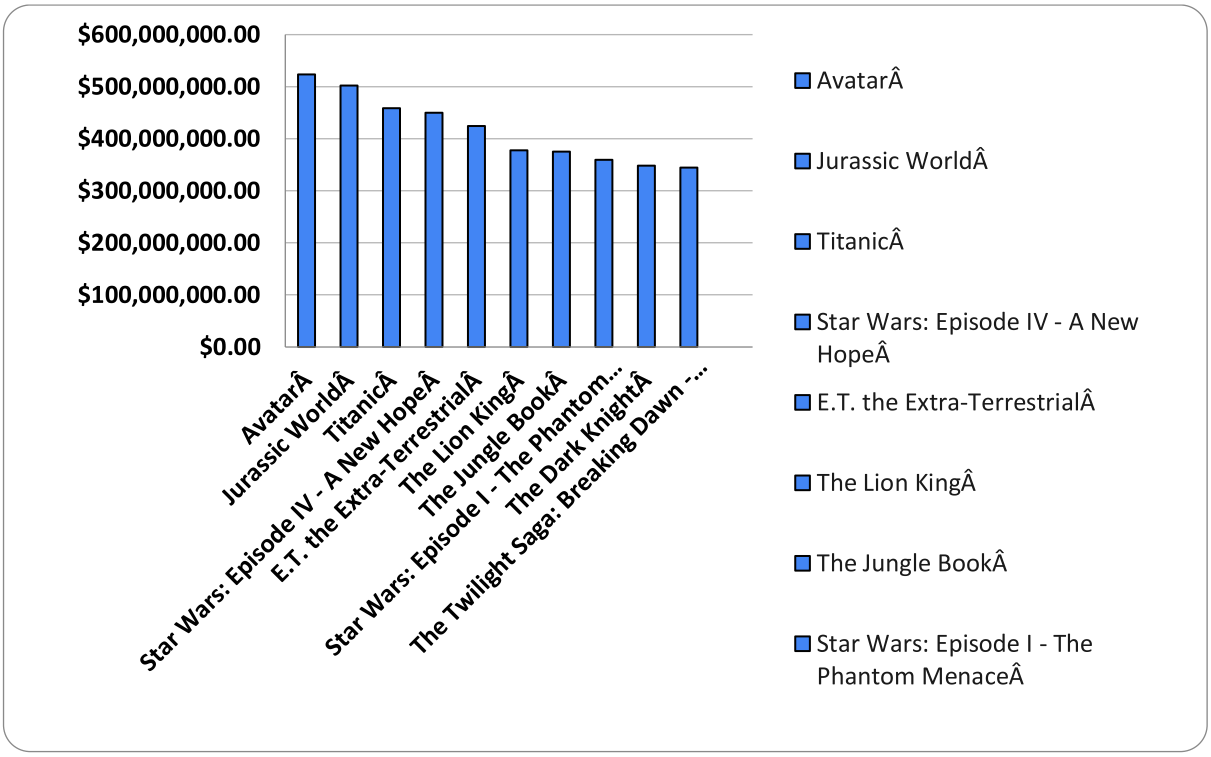
=INDEX(B2:B3849, MATCH(1,IF(D2:D3849=G11, 1),0)) to get title of movie.

Output:



5) PROFIT ANALYSIS:-

Movies	Profits in Millions
AvatarÂ	523505847
Jurassic WorldÂ	502177271
TitanicÂ	458672302
Star Wars: Episode IV - A N	449935665
E.T. the Extra-TerrestrialÂ	424449459
The Lion KingÂ	377783777
The Jungle BookÂ	375290282
Star Wars: Episode I - The	359544677
The Dark KnightÂ	348316061
The Twilight Saga: Breakin	344597846



Movies with highest profit margin is Avatar.

With the help of this project, I have gained valuable experience for data analysis using statistical knowledge and excel's data visualization. Through this, I have learnt to apply my data analysis skills in solving real life problems.

LINK for cleaned Data Set:

Click here to see Excel file