IMDB MOVIE ANALYSIS

DESCRIPTION:

IMDB means Internet Movie database. It is qiuet similar to Netflix where we get related information about movies, television shows, directors, biographies, plot summaries, reviews, video games and related information. Later it became a website and become a popular online platform where we can find information related to movies.

PROBLEM STATEMENT:

The dataset provided is related to IMDB Movies. A potential problem to investigate could be: "What factors influence the success of a movie on IMDB?" Here, success can be defined by high IMDB ratings. The impact of this problem is significant for movie producers, directors, and investors who want to understand what makes a movie successful to make informed decisions in their future projects.

DATA CLEANING:

This step involves preprocessing the data to make it suitable for analysis. It includes handling missing values, removing duplicates, converting data types if necessary, and possibly feature engineering.

DATA ANALYSIS:

Here, you'll explore the data to understand the relationships between different variables. You might look at the correlation between movie ratings and other factors like genre, director, budget, etc. You might also want to consider the year of release, the actors involved, and other relevant factors.

DATA ANALYTICS TASKS:

You are required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

A. Movie Genre Analysis: Analyse 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.

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

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

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

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

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

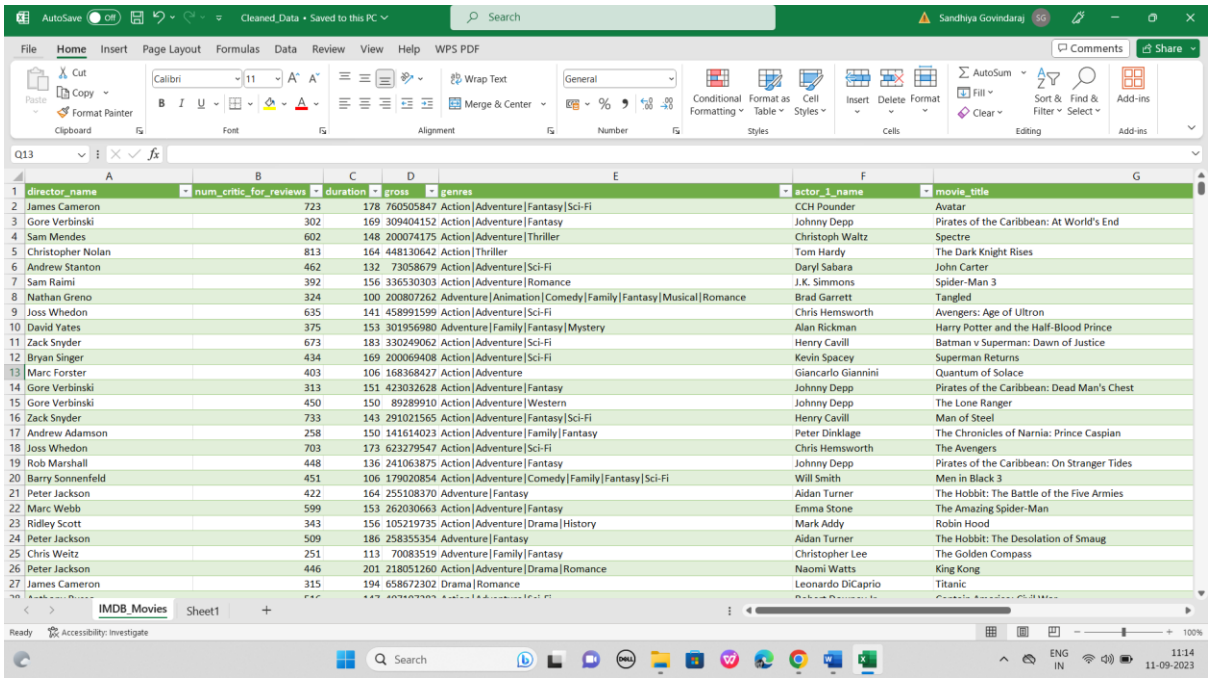
Task: Identify the top directors based on their average IMDB score and analyse their contribution to the success of movies using percentile calculations.

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

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

Tech-Stack Used: Microsoft-Excel

Before Analyses we want to clean the data. We need to remove unwanted columns and Duplicate/Null values.



director_name	num_critic_for_reviews	duration	gross	genres	actor_1_name	movie_title
James Cameron	723	178	760505847	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	Avatar
Gore Verbinski	302	169	309404152	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End
Sam Mendes	602	148	200074175	Action|Adventure|Thriller	Christoph Waltz	Spectre
Christopher Nolan	813	164	448130642	Action|Thriller	Tom Hardy	The Dark Knight Rises
Andrew Stanton	462	132	73058679	Action|Adventure|Sci-Fi	Daryl Sabara	John Carter
Sam Raimi	392	156	336530303	Action|Adventure|Romance	J.K. Simmons	Spider-Man 3
Nathan Greno	324	100	200807262	Adventure|Animation|Comedy|Family|Fantasy|Musical|Romance	Brad Garrett	Tangled
Joss Whedon	635	141	458991599	Action|Adventure|Sci-Fi	Chris Hemsworth	Avengers: Age of Ultron
David Yates	375	153	301956980	Adventure|Family|Fantasy|Mystery	Alan Rickman	Harry Potter and the Half-Blood Prince
Zack Snyder	673	183	330249062	Action|Adventure|Sci-Fi	Henry Cavill	Batman v Superman: Dawn of Justice
Bryan Singer	434	169	200069408	Action|Adventure|Sci-Fi	Kevin Spacey	Superman Returns
Marc Forster	403	106	168368427	Action|Adventure	Giancarlo Giannini	Quantum of Solace
Gore Verbinski	313	151	423032628	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: Dead Man's Chest
Gore Verbinski	450	150	89289910	Action|Adventure|Western	Johnny Depp	The Lone Ranger
Zack Snyder	733	143	291021565	Action|Adventure|Fantasy|Sci-Fi	Henry Cavill	Man of Steel
Andrew Adamson	258	150	141614023	Action|Adventure|Family|Fantasy	Peter Dinklage	The Chronicles of Narnia: Prince Caspian
Joss Whedon	703	173	623279547	Action|Adventure|Sci-Fi	Chris Hemsworth	The Avengers
Rob Marshall	448	136	241063875	Action|Adventure|Fantasy	Johnny Depp	Pirates of the Caribbean: On Stranger Tides
Barry Sonnenfeld	451	106	179020854	Action|Adventure|Comedy|Family|Fantasy|Sci-Fi	Will Smith	Men in Black 3
Peter Jackson	422	164	255108370	Adventure|Fantasy	Aidan Turner	The Hobbit: The Battle of the Five Armies
Marc Webb	599	153	262030663	Action|Adventure|Fantasy	Emma Stone	The Amazing Spider-Man
Ridley Scott	343	156	105219735	Action|Adventure|Drama|History	Mark Addy	Robin Hood
Peter Jackson	509	186	258355354	Adventure|Fantasy	Aidan Turner	The Hobbit: The Desolation of Smaug
Chris Weitz	251	113	70083519	Adventure|Family|Fantasy	Christopher Lee	The Golden Compass
Peter Jackson	446	201	218051260	Action|Adventure|Drama|Romance	Naomi Watts	King Kong
James Cameron	315	194	658672302	Drama|Romance	Leonardo DiCaprio	Titanic

I have provided the full view of excel sheet on google drive.

1. **Movie Genre Analysis:**
    - Select genre column from cleaned_data.
    - Separate genre column to provide comman genres report
    - After finding unique genre for each find mean, median, mode, range, variance and standard deviation with imdb_score column.

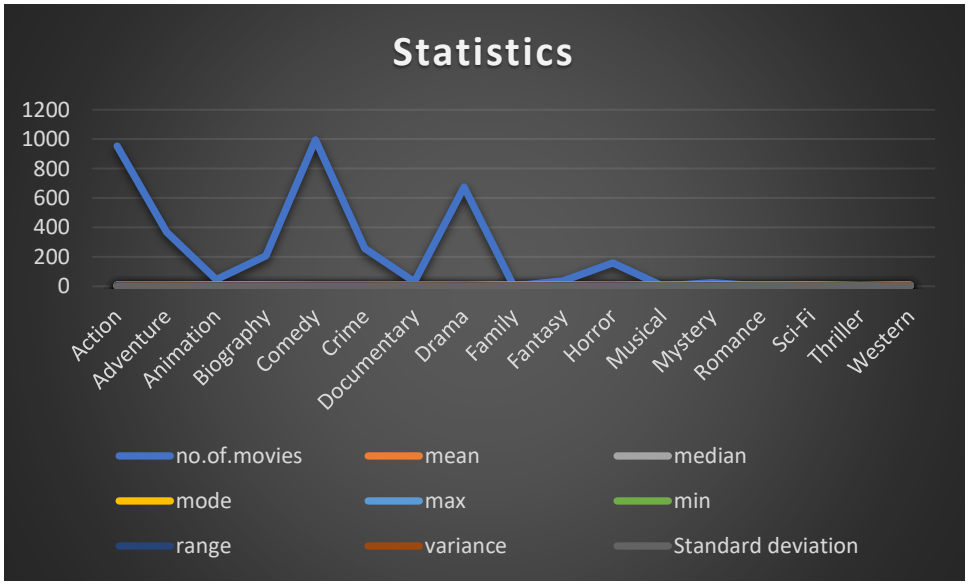
Functions used:

To count no.of movies use COUNTIF function.

Mean	= AVERAGE (range,criteria)
Median	= MEDIAN (range,criteria)
Mode	= MODE (range,criteria)
Range	= max-min
Variance	= VAR (range,criteria)
Standard deviation	= STDEV (range,criteria)

Most Common genres and descriptive statistics with imdb_score

genres	no.of.movies	mean	median	mode	max	min	range	variance	Standard deviation
Action	954	6.290671	6.3	6.1	9	2.1	6.9	1.069388	1.034112284
Adventure	368	6.560326	6.7	7.3	8.6	2.3	6.3	1.258095	1.121648214
Animation	45	6.74	7	7.1	8	4.5	3.5	0.942455	0.970800981
Biography	206	7.160194	7.2	7	8.9	4.5	4.4	0.480944	0.693501516
Comedy	995	6.169849	6.3	6.4	8.8	1.9	6.9	1.066052	1.032497839
Crime	257	6.933852	7	7.4	9.3	3.3	6	0.767092	0.875837738
Documentary	30	6.806667	7.4	7.5	8.5	1.6	6.9	2.723402	1.650273401
Drama	675	6.826815	6.9	7.3	8.8	2.1	6.7	0.822559	0.906950291
Family	3	6.5	5.9	#N/A	7.9	5.7	2.2	1.48	1.216552506
Fantasy	37	6.281081	6.5	6.8	7.9	4.3	3.6	0.799354	0.894066191
Horror	159	5.842138	5.9	5.9	8.5	2.3	6.2	1.059795	1.029463646
Musical	2	6.75	6.75	#N/A	7.2	6.3	0.9	0.405	0.636396103
Mystery	23	6.652174	6.7	7.1	8.5	3.3	5.2	1.193518	1.092482396
Romance	2	6.65	6.65	#N/A	7.1	6.2	0.9	0.405	0.636396103
Sci-Fi	8	6.5875	6.35	#N/A	8.2	5	3.2	1.064107	1.031555691
Thriller	1	4.8	4.8	4.8	4.8	4.8	0	0	0
Western	3	6.766667	7.3	#N/A	8.9	4.1	4.8	5.973333	2.444040371



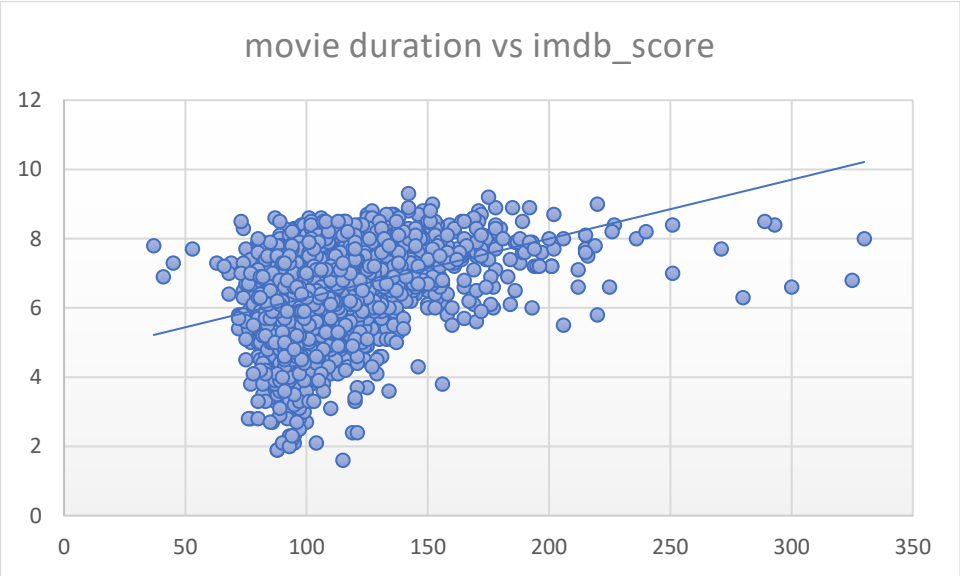
Most comman genre is Comedy.

Three are 17 genres.

2. Movie Duration Analysis:

- Select duration column and imdb_score for analysis from cleaned_data
- Find duration statistics with functions of mean, median and standard deviation.
- Scatter plot to visualize the relationship between movie durations and imdb_score.
- Use tredline for direction.

movie durations	
mean	110.2153
median	106
Standard deviation	22.72277



Function used:

Mean = AVERAGE (range, criteria)

Median = MEDIAN (range, criteria)

Standard Deviation = STDEV(range, criteria)

3. Language Analysis:

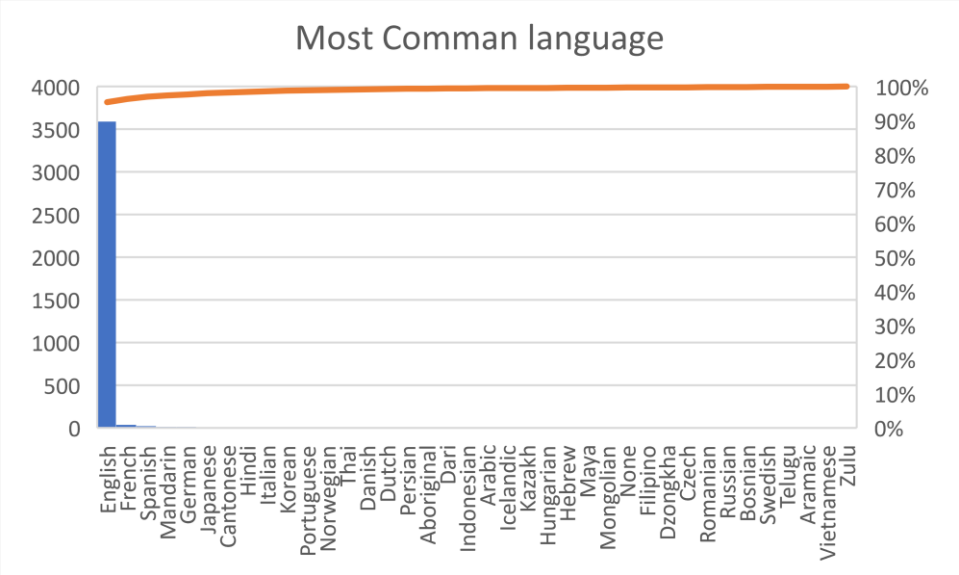
- Select language column for unique languages from cleaned_data
- Calculate the statistics of mean, median and standard deviation for each languages.

Unique languages	no.of.movies	mean	median	standard deviation
Aboriginal	2	6.95	6.95	0.777817459
Arabic	1	7.2	7.2	0
Aramaic	1	7.1	7.1	0
Bosnian	1	4.3	4.3	0
Cantonese	8	7.2375	7.3	0.440575922
Czech	1	7.4	7.4	0
Danish	3	7.9	8.1	0.529150262
Dari	2	7.5	7.5	0.141421356
Dutch	3	7.566667	7.8	0.404145188
Dzongkha	1	7.5	7.5	0
English	3593	6.426273	6.5	1.049758669
Filipino	1	6.7	6.7	0
French	36	7.297222	7.25	0.565425812
German	13	7.692308	7.7	0.640912811
Hebrew	1	8	8	0
Hindi	8	7.175	7.3	0.761108215
Hungarian	1	7.1	7.1	0
Icelandic	1	6.9	6.9	0
Indonesian	2	7.9	7.9	0.424264069
Italian	7	7.185714	7	1.155318962
Japanese	12	7.625	7.8	0.899621132
Kazakh	1	6	6	0
Korean	5	7.7	7.7	0.570087713
Mandarin	14	7.021429	7.25	0.765786244
Maya	1	7.8	7.8	0
Mongolian	1	7.3	7.3	0
None	1	8.5	8.5	0
Norwegian	4	7.15	7.3	0.574456265
Persian	3	8.133333	8.4	0.550757055
Portuguese	5	7.76	8	0.978774744
Romanian	1	7.9	7.9	0
Russian	1	6.5	6.5	0
Spanish	23	7.082609	7.2	0.860577065
Swedish	1	7.6	7.6	0
Telugu	1	8.4	8.4	0
Thai	3	6.633333	6.6	0.450924975
Vietnamese	1	7.4	7.4	0
Zulu	1	7.3	7.3	0

Function used:

COUNTIF to count no.of movies for each languages.

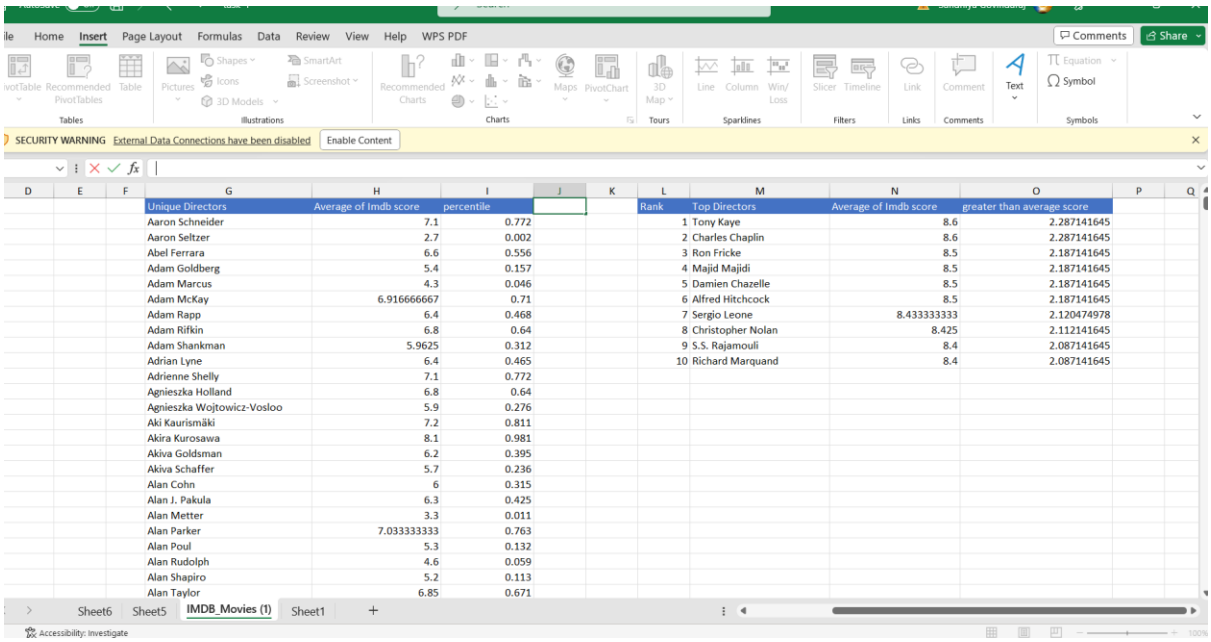
Mean = AVERAGE (range, criteria)
Median = MEDIAN (range, criteria)
Standard Deviation = STDEV (range, criteria)



Most Comman Language is English.

4. Director Analysis:

- Select director column and imdb_score column from cleaned_data.
- Separate unique directors and find each average score with imdb_score using AVERAGE(range,criteria) function.
- Find percentile for each unique directors with PERCENTRANK.EXC function.
- Select top directors with highest average score and compare with overall distribution of scores.



Unique Directors	Average of Imdb score	percentile	Rank	Top Directors	Average of Imdb score	greater than average score
Aaron Schneider	7.1	0.772	1	Tony Kaye	8.6	2.287141645
Aaron Seltzer	2.7	0.002	2	Charles Chaplin	8.6	2.287141645
Abel Ferrara	6.6	0.556	3	Ron Fricke	8.5	2.187141645
Adam Goldberg	5.4	0.157	4	Majid Majidi	8.5	2.187141645
Adam Marcus	4.3	0.046	5	Damien Chazelle	8.5	2.187141645
Adam McKay	6.916666667	0.71	6	Alfred Hitchcock	8.5	2.187141645
Adam Rapp	6.4	0.468	7	Sergio Leone	8.433333333	2.120474978
Adam Rifkin	6.8	0.64	8	Christopher Nolan	8.425	2.112141645
Adam Shankman	5.9625	0.312	9	S.S. Rajamouli	8.4	2.087141645
Adrian Lyne	6.4	0.465	10	Richard Marquand	8.4	2.087141645
Adrienne Shelly	7.1	0.772				
Agnieszka Holland	6.8	0.64				
Agnieszka Wojtowicz-Vosloo	5.9	0.276				
Aki Kaurismäki	7.2	0.811				
Akira Kurosawa	8.1	0.981				
Akiva Goldsman	6.2	0.395				
Akiva Schaffer	5.7	0.236				
Alan Cohn	6	0.315				
Alan J. Pakula	6.3	0.425				
Alan Metter	3.3	0.011				
Alan Parker	7.033333333	0.763				
Alan Poul	5.3	0.132				
Alan Rudolph	4.6	0.059				
Alan Shapiro	5.2	0.113				
Alan Taylor	6.85	0.671				

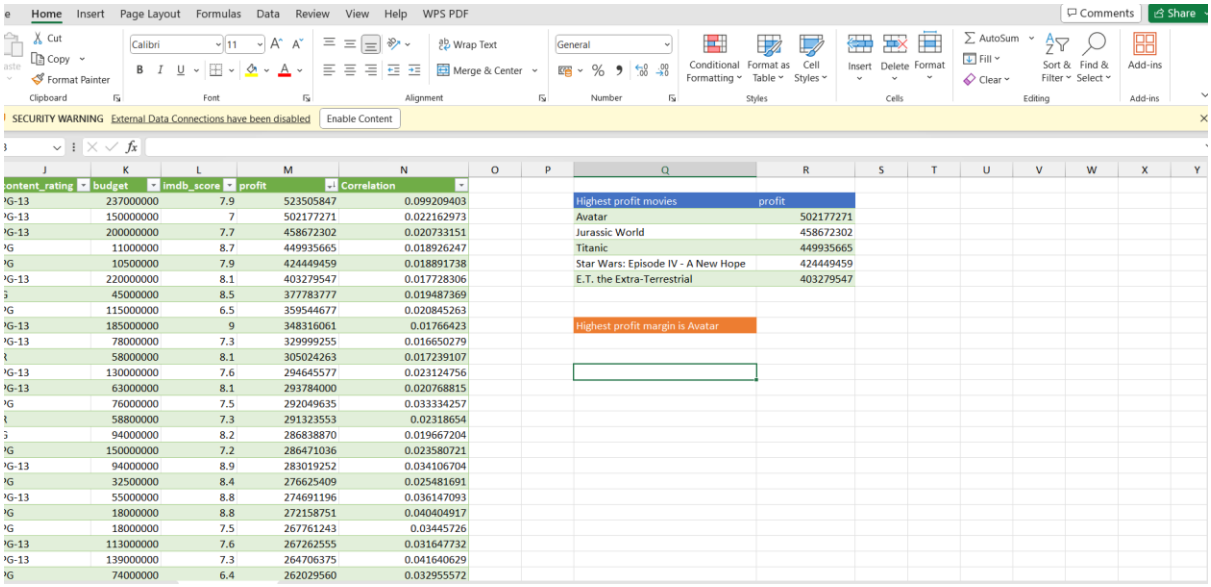
To view full worksheet follow the link provided on google drive.

5. Budget Analysis:

- Select budget and gross column to find profit. Difference between gross and budget is the result of profit.
- Find correlation between budget and profit
- Select top movies with highest profit
- Find highest profit margin using MAX(range) function.

Highest profit movies	profit
Avatar	502177271
Jurassic World	458672302
Titanic	449935665
Star Wars: Episode IV - A New Hope	424449459
E.T. the Extra-Terrestrial	403279547

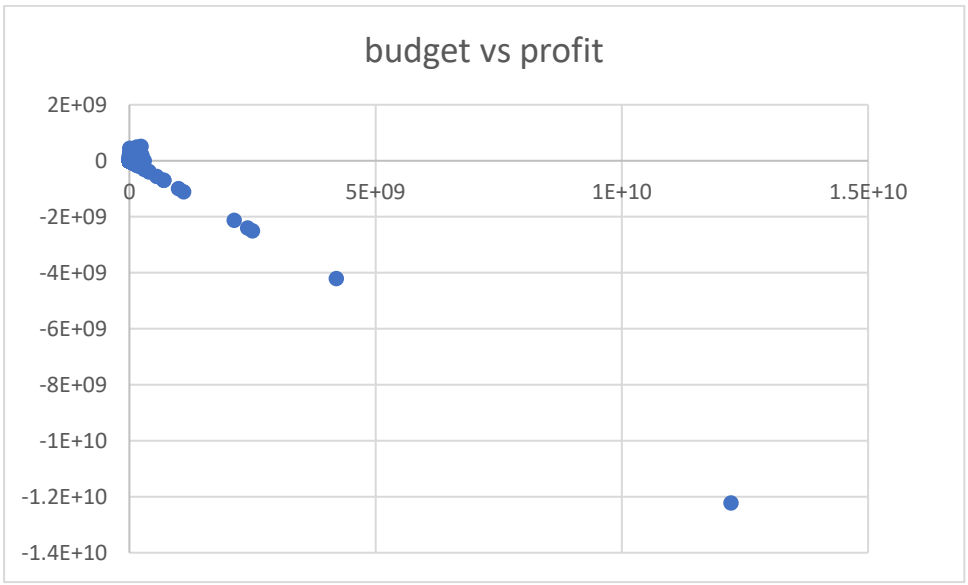
Highest profit movie is Avatar.



content_rating	budget	imdb_score	profit	Correlation
PG-13	237000000	7.9	523505847	0.099209403
PG-13	150000000	7	502177271	0.022162973
PG-13	200000000	7.7	458672302	0.020733151
PG	11000000	8.7	449935665	0.018926247
PG	10500000	7.9	424449459	0.018891738
PG-13	220000000	8.1	403279547	0.017728306
G	45000000	8.5	377783777	0.019487369
PG	115000000	6.5	359544677	0.020845263
PG-13	185000000	9	348316061	0.01766423
PG-13	78000000	7.3	329999255	0.016650279
R	58000000	8.1	305024263	0.017239107
PG-13	130000000	7.6	294645577	0.023124756
PG-13	63000000	8.1	293784000	0.020768815
PG	76000000	7.5	292049635	0.033334257
R	58800000	7.3	291323553	0.02318654
G	94000000	8.2	286838870	0.019667204
PG	150000000	7.2	286471036	0.023580721
PG-13	94000000	8.9	283019252	0.034106704
PG	32500000	8.4	276625409	0.025481691
PG-13	55000000	8.8	274691196	0.036147093
PG	18000000	8.8	272158751	0.040404917
PG	18000000	7.5	267761243	0.03445726
PG-13	113000000	7.6	267262555	0.031647732
PG-13	139000000	7.3	264706375	0.041640629
PG	74000000	6.4	262029560	0.032955572

Highest profit movies	profit
Avatar	502177271
Jurassic World	458672302
Titanic	449935665
Star Wars: Episode IV - A New Hope	424449459
E.T. the Extra-Terrestrial	403279547

Highest profit margin is Avatar



To view full worksheet you can refer to excel sheet on google drive link

I have provided all my excel sheets on google drive link Here is the link for your reference.

Google drive link :

https://drive.google.com/drive/folders/1dzv2ZfEeK-QxF_wzTnZ0WL9Wm6cLHx2R?usp=sharing