IMDB Movie Analysis

Final Project-1

Project Description:

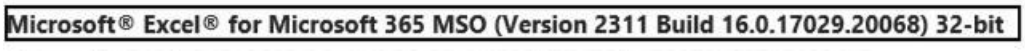
The data supplied is related to IMDB movies. One such problem to study is "what factors influence the success of a movie on IMDB?". In this context, success is determined by high IMDB ratings. This issue has a huge impact on movie producers, directors, and financiers who want to understand what makes a film successful so that they can make informed judgements about future projects.

Approach:

- First, I downloaded the dataset and made a copy of it. Then, I read over the problem statement and tasks that I needed to complete to gain insights. The given dataset was imported into excel. After importing the data, I cleaned it up to make it acceptable for analysis.
- It includes removing unneeded data for specific tasks- The initial dataset had 28 columns and 5044 rows. All columns are not required to obtain insights. So, I removed unnecessary columns and decreased to 8 columns. I also added 2 new columns, one for genre(separated), which is divided using flash fill and another for profit. I now have 10 columns.
- Then after I deleted duplicate rows using the Remove Duplicates from data tab.
- After that, I used 'find & select' followed by 'go to special 'to handle all empty rows before selecting the 'blanks' option. This operation highlighted all empty rows. After words, I used the shortcut 'ctrl + - 'to choose the full rows option. This approach resulted in the removal of all empty entries.
- from the dataset. Finally, now have a cleaned dataset with **10 columns** and **3790 rows.**
- With the cleaned dataset, I began looking for insights into certain jobs.to do the analysis, I utilized excel 365 and pivot tables, graphs, charts, and other

tools to visualize data and create patterns. After that I draw conclusions and offered recommendations based on the patterns I discovered, and
I also used Microsoft word to create a complete project report.

Tech-Stack Used:



Working on spreadsheets in real time with auto save and update is possible when utilizing excel 365.
As with excel 365, it provides several strong visualization features that can aid in the analysis and comprehension of data. I utilized capabilities like pivot tables, graphs, charts, and formatting options.
And MS. Word to write a project report. A loom video explaining whole project.

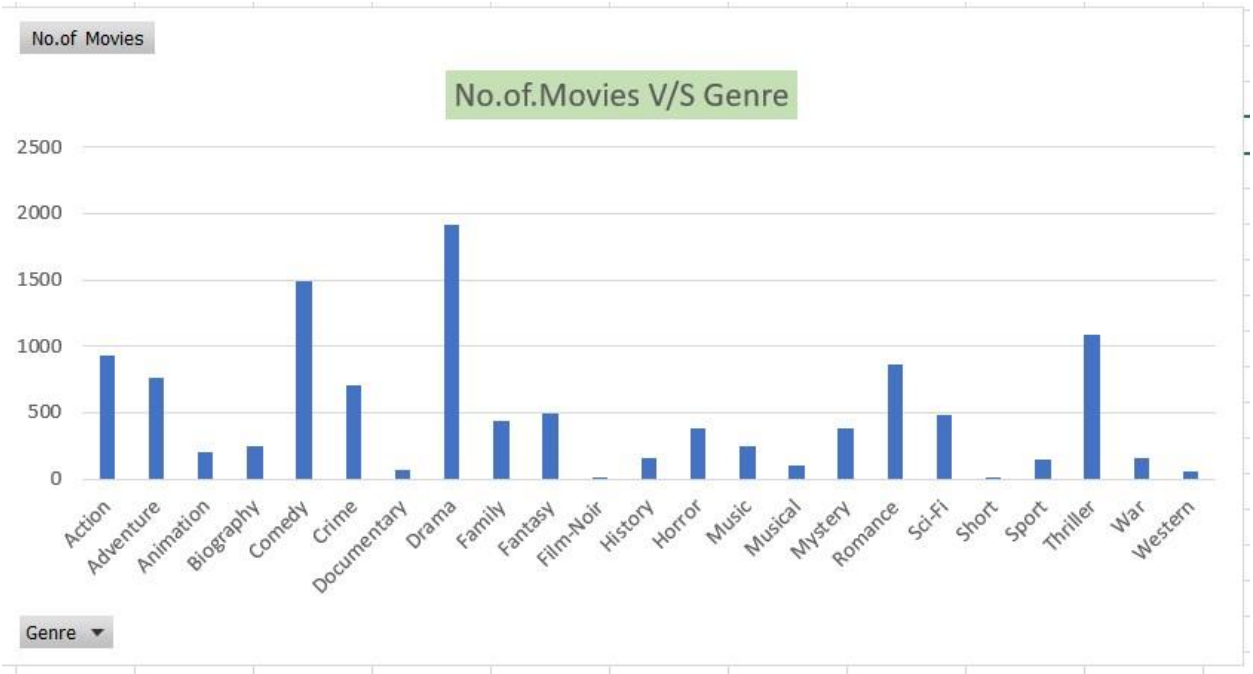
Insights:

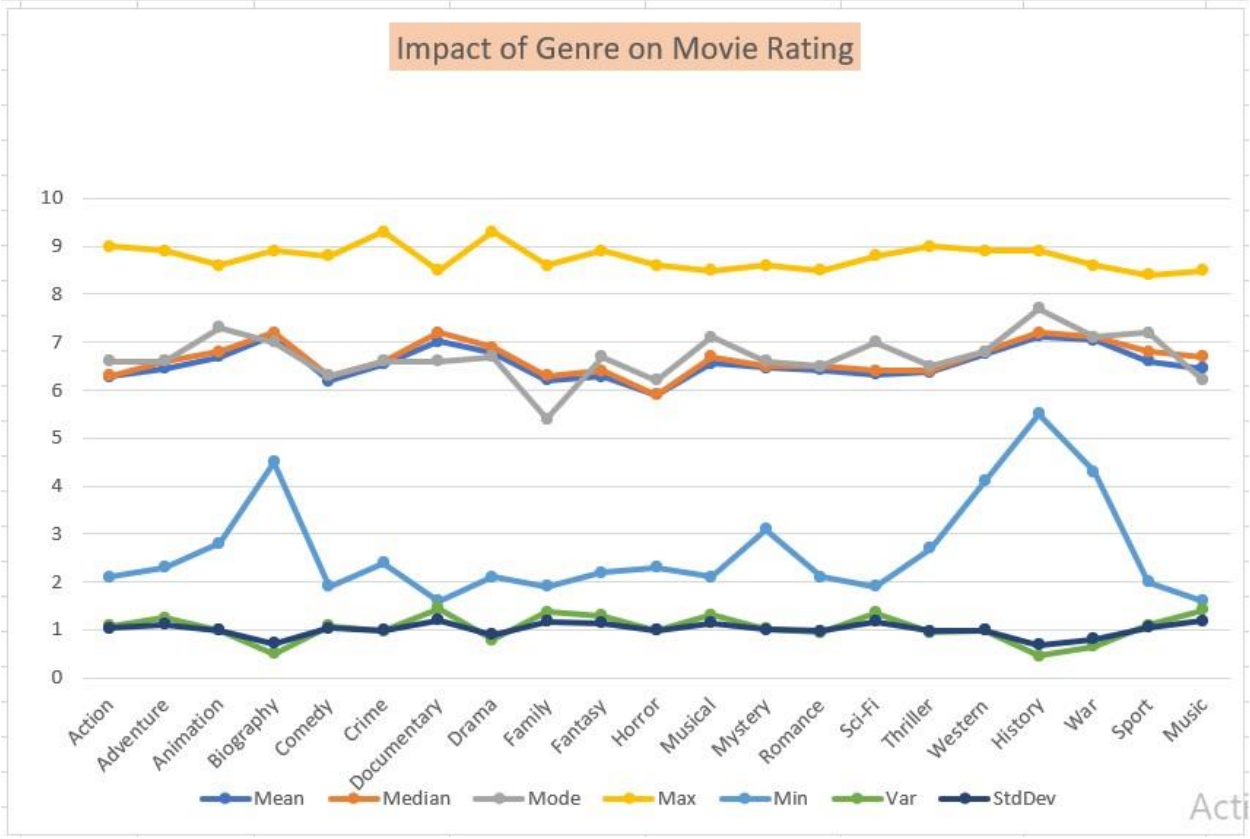
A. Movie Genre Analysis: Analyse the distribution of movie genres and their impact on IMDB score.

Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics.

The most prevalent movie genres are mentioned below along with descriptive statistics.

Genre	No.of Movies	Mean	Median	Mode	Max	Min	Var	StdDev
Action	935	6.285989305	6.3	6.6	9	2.1	1.078186788	1.038357736
Adventure	766	6.454960836	6.6	6.6	8.9	2.3	1.247524378	1.116926308
Animation	197	6.700507614	6.8	7.3	8.6	2.8	0.987295659	0.993627525
Biography	242	7.140082645	7.2	7	8.9	4.5	0.504237338	0.71009671
Comedy	1492	6.183310992	6.3	6.3	8.8	1.9	1.081431552	1.039919012
Crime	703	6.545661451	6.6	6.6	9.3	2.4	0.97143058	0.98561178
Documentary	67	7.011940299	7.2	6.6	8.5	1.6	1.439855269	1.199939694
Drama	1914	6.787774295	6.9	6.7	9.3	2.1	0.795417626	0.891861887
Family	442	6.202262443	6.3	5.4	8.6	1.9	1.3670697	1.169217559
Fantasy	496	6.285080645	6.4	6.7	8.9	2.2	1.30054464	1.140414241
Horror	379	5.903957784	5.9	6.2	8.6	2.3	0.982127152	0.991023285
Musical	102	6.550980392	6.7	7.1	8.5	2.1	1.307672297	1.143535
Mystery	377	6.469496021	6.5	6.6	8.6	3.1	1.014838309	1.007391835
Romance	867	6.425490196	6.5	6.5	8.5	2.1	0.938321786	0.968670112
Sci-Fi	484	6.327272727	6.4	7	8.8	1.9	1.362318841	1.16718415
Thriller	1087	6.372309108	6.4	6.5	9	2.7	0.939112803	0.969078327
Western	60	6.756666667	6.8	6.8	8.9	4.1	0.982158192	0.991038946
History	153	7.122875817	7.2	7.7	8.9	5.5	0.460196938	0.678378168
War	159	7.048427673	7.1	7.1	8.6	4.3	0.652386753	0.80770462
Sport	147	6.601360544	6.8	7.2	8.4	2	1.09876526	1.048220043
Music	247	6.456680162	6.7	6.2	8.5	1.6	1.413359666	1.188848041
Short	2	6.8	6.8	#N/A	7.1	6.5	0.18	0.424264069
Film-Noir	1	7.7	7.7	#N/A	7.7	7.7	#DIV/0!	#DIV/0!





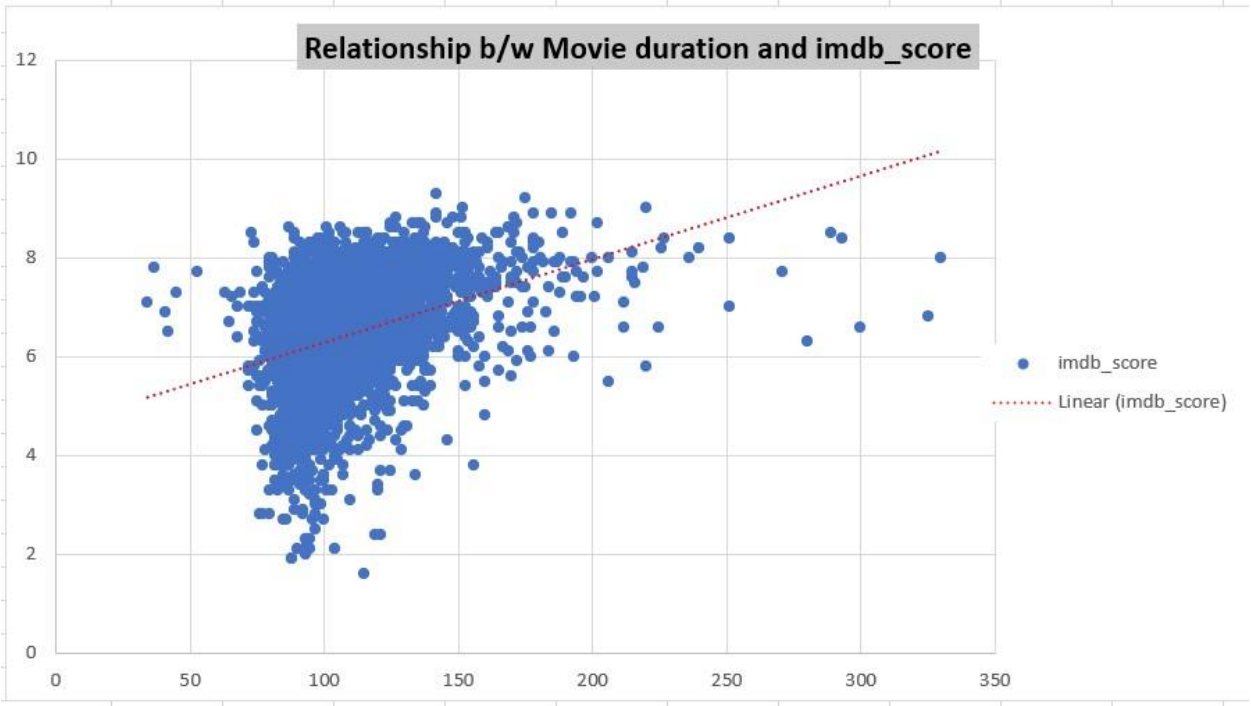
Insights:

- **Drama** is the most common movie genre in films, followed by **comedy, thriller, action, and romance**.
- The genre with the fewest number of films were film-noir (1), and short (2).
- **Biography, history, documentary, and war** have the highest mean,
- median, impact on movie ratings.
- Crime (9.3), drama (9.3), action (9), and thriller (9) have received the maximum rating among the other genres. Whereas documentary and music received the minimum ratings.

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

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

operations	values
Mean	109.8029
Median	105
Mode	101
Variance	517.8906
Std Deviation	22.75721



Insights:

the above table displays descriptive statistics regarding movie duration, including mean, median, and standard deviation.

- The movie duration data has no outliers. It's because the mean and median differ so little. The standard deviation is 22.75721, which is large and shows that the data values deviate significantly from mean.

- By displaying relationship between movie duration and IMDB score, it is obvious that the trendline in the scatterplot is slightly upward, indicating a moderate positive link between duration and IMDB score.
- The above trendline, maximum movie duration ranges from 75min to 55min, with IMDB score ranging from 5 to 8 ratings.
- The movie with extended durations between 150 min to 330 min have IMDB scores ranging from 6 to 9 ratings.
- There is a strong impact on durations between 80 to 190 minutes, with average IMDB scores exceeding 8 ratings.

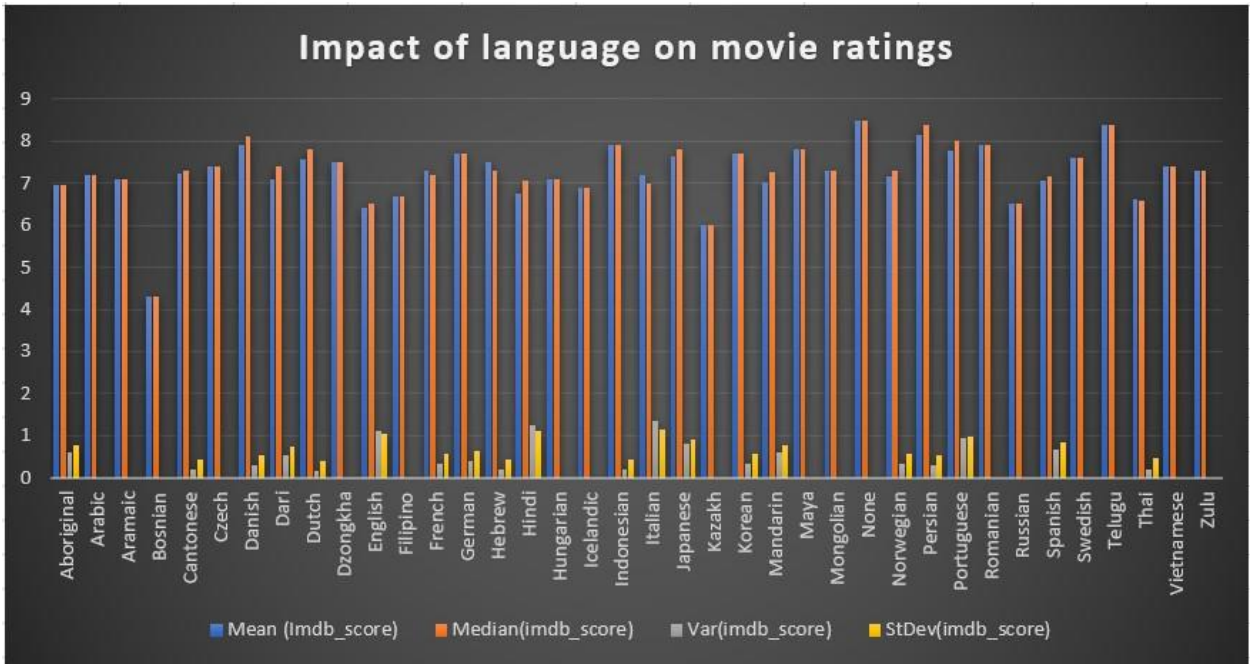
C. Language Analysis: situation: examine the distribution of movies based on their languages.

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

descriptive statistics for language analysis

Languages	No.of.Movies	Mean (Imdb_score)	Median(imdb_score)	Var(imdb_score)	StDev(imdb_score)
Aboriginal	2	6.95	6.95	0.605	0.777817459
Arabic	1	7.2	7.2	#DIV/0!	#DIV/0!
Aramaic	1	7.1	7.1	#DIV/0!	#DIV/0!
Bosnian	1	4.3	4.3	#DIV/0!	#DIV/0!
Cantonese	8	7.2375	7.3	0.194107143	0.440575922
Czech	1	7.4	7.4	#DIV/0!	#DIV/0!
Danish	3	7.9	8.1	0.28	0.529150262
Dari	2	7.08125	7.4	0.536291667	0.732319375
Dutch	3	7.566666667	7.8	0.163333333	0.404145188
Dzongkha	1	7.5	7.5	#DIV/0!	#DIV/0!
English	3606	6.421436495	6.5	1.107753941	1.052498903
Filipino	1	6.7	6.7	#DIV/0!	#DIV/0!
French	37	7.286486486	7.2	0.31509009	0.561328861
German	13	7.692307692	7.7	0.410769231	0.640912811
Hebrew	3	7.5	7.3	0.19	0.435889894
Hindi	10	6.76	7.05	1.236	1.111755369
Hungarian	1	7.1	7.1	#DIV/0!	#DIV/0!
Icelandic	1	6.9	6.9	#DIV/0!	#DIV/0!
Indonesian	2	7.9	7.9	0.18	0.424264069
Italian	7	7.185714286	7	1.334761905	1.155318962
Japanese	12	7.625	7.8	0.809318182	0.899621132
Kazakh	1	6	6	#DIV/0!	#DIV/0!
Korean	5	7.7	7.7	0.325	0.570087713

Korean	5	7.7	7.7	0.325	0.570087713
Mandarin	14	7.021428571	7.25	0.586428571	0.765786244
Maya	1	7.8	7.8	#DIV/0!	#DIV/0!
Mongolian	1	7.3	7.3	#DIV/0!	#DIV/0!
None	1	8.5	8.5	#DIV/0!	#DIV/0!
Norwegian	4	7.15	7.3	0.33	0.574456265
Persian	3	8.133333333	8.4	0.303333333	0.550757055
Portuguese	5	7.76	8	0.958	0.978774744
Romanian	1	7.9	7.9	#DIV/0!	#DIV/0!
Russian	1	6.5	6.5	#DIV/0!	#DIV/0!
Spanish	26	7.05	7.15	0.6826	0.826196103
Swedish	1	7.6	7.6	#DIV/0!	#DIV/0!
Telugu	1	8.4	8.4	#DIV/0!	#DIV/0!
Thai	3	6.633333333	6.6	0.203333333	0.450924975
Vietnamese	1	7.4	7.4	#DIV/0!	#DIV/0!
Zulu	1	7.3	7.3	#DIV/0!	#DIV/0!



Insights:

- **English** is the most used common language in movies (3606), followed by **French (37), Spanish (26), Mandarin (14), German (13), Japan (12), and Hindi (10).** The remaining languages have fewer than 10 movies.
- **Telugu and Persian** languages have the highest mean (8.4,8.1) and median (8.4,8.4) IMDB score, while **Bosnian** has the lowest mean (4.3) and median (4.3).

- The standard deviation of IMDB score is less than the mean of IMDB score for each language and is positive.

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.

Approach for this task- I calculated the average IMDB score for each director. I also used percentile function to determine which directors had the highest rating. Following this, I utilised a number filter to determine the top 12 directors based on percentile data.

Top 12 directors according to their average imdb scores with percentile:

director_name	No.of Movi	Average(imdb_scor	Percent
Tony Kaye	1	8.6	0.998
Charles Chaplin	1	8.6	0.998
Alfred Hitchcock	1	8.5	0.996
Ron Fricke	1	8.5	0.996
Damien Chazelle	1	8.5	0.996
Majid Majidi	1	8.5	0.996
Sergio Leone	3	8.433333333	0.996
Christopher Nolan	8	8.425	0.995
S.S. Rajamouli	1	8.4	0.993
Richard Marquand	1	8.4	0.993
Asghar Farhadi	1	8.4	0.993
Marius A. Markevicius	1	8.4	0.993

Insight:

- Here are the top 12 directors based on their average IMDB-score with percentile. The highest percentile is 0.998 for directors **Tony Kaye** and **Charles Chaplin.**
- The directors whose films have received an average an average IMDB score of at least 7.5 are regarded to be top directors.

E. Budget analysis: explore the relationship between movie budget and their financial success.

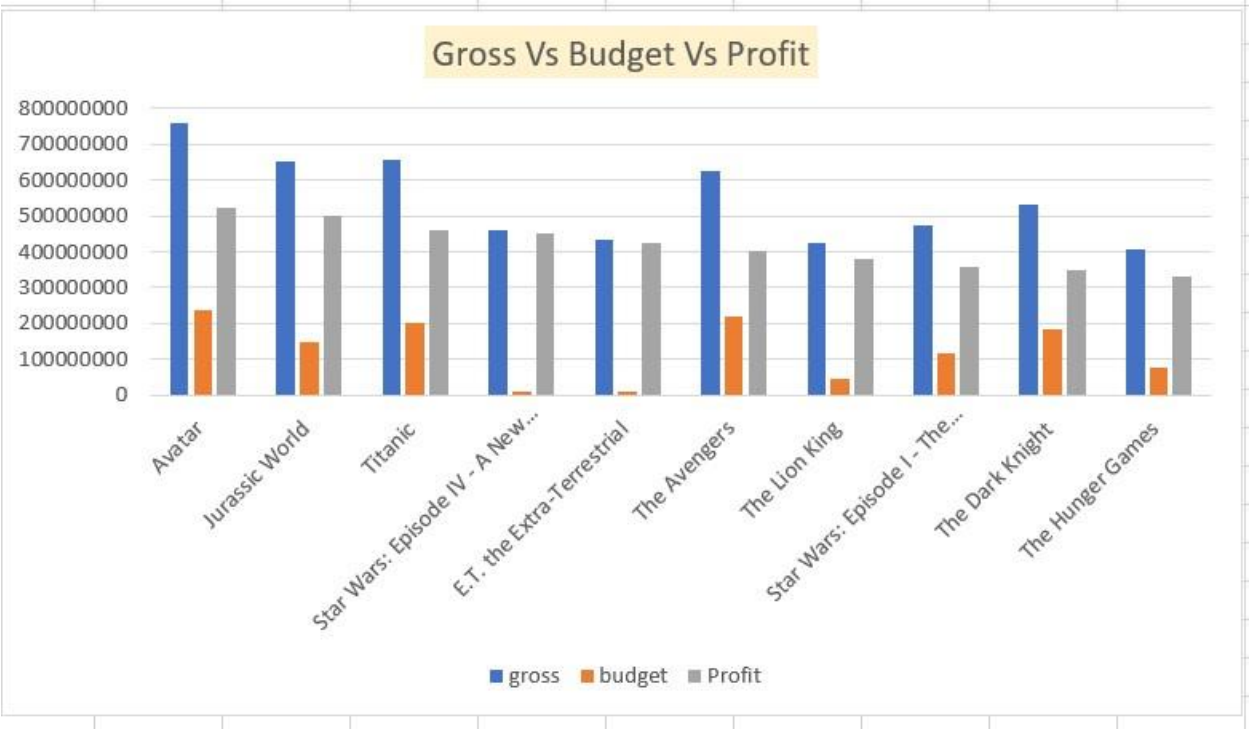
Task: analyse the correlation between movie budgets and gross earnings, identify the movies with the highest profit margin.

correlation coefficient:	0.0966483

Movie_Title	gross	budget	Profit
Avatar	760505847	237000000	523505847
Jurassic World	652177271	150000000	502177271
Titanic	658672302	200000000	458672302
Star Wars: Episode IV - A New Hope	460935665	11000000	449935665
E.T. the Extra-Terrestrial	434949459	10500000	424449459
The Avengers	623279547	220000000	403279547
The Lion King	422783777	45000000	377783777
Star Wars: Episode I - The Phantom Menace	474544677	115000000	359544677
The Dark Knight	533316061	185000000	348316061
The Hunger Games	407999255	78000000	329999255

Top 10 movies with highest profit margin are shown in the graph.

The relation between gross, budget, and profit:



Insights:

- The correlation coefficient of **0.096514** is near to zero, implying a modest tendency for movies with larger budgets to have somewhat higher gross earnings. As a result, the correlation between budget and profit is weak.
- **Avatar** has the greatest profit margin of **523505847 million dollars**, followed by **Jurassic world**, both of which are **very high budget films.**
- When compared to the top ten movies, **Star Wars Episode IV: A New Hope, E.T. the Extra-Terrestrial**, and The Lion King make a lot of profit. Which are **low budget movies.**

Result:

Findings and recommendations:

- Although drama is the most prevalent genre, I would recommend making a film containing the biography, history, documentary, and war genres. It's because these genres have the highest average IMDB rating. These genres can have a significant impact on movie ratings.
- I discovered that movies with a duration of 80 to 200 minutes have the most influence. If the movie storyline, cast, and crew are good, and the duration is between 80-200 minutes or although duration is more, the IMDB score can be greater than 7.5, and the movie's success rate is high.
- The most used language in movies is English, followed by French, Spanish, mandarin, German, Japan, and Hindi. Although movies made in other languages can have significant impact on IMDB score, if they are dubbed into most widely used languages or have subtitles added to these movies in popular languages. Profits are also possibles.
- Tony Koye and Charlie Chaplin are the two filmmakers with the highest average IMDB movie rating. I advise that movie makers get inspiration from the top ten directors. In addition, film production companies can collaborate with these top directors.
- The most profitable films are between $100 and $300 million. To reduce the risk of loss, film production business should maintain their budgets within the range.

impact:

  - By working on this project, I learned how to handle big dataset, clean data, and use editing functions like Find & Select, Sort & Filter, as well as the flash fill option, percentile, correlation, and descriptive statistics in excel.
  - By working as a data analytics, I learned to study audience preferences, and identify target audiences, as well as trends in the film industry based on insights. I also learned how to record a loom video to explain this project.

    EXCEL Sheet link – working sheet P5-IMDB MOVIE ANALYSIS.xlsx

Video Presentation -
https://www.loom.com/share/2f7fdc11cfdb49f5ad8e9252db6f0686?sid=86ba7e0e-68e5-4eb5-9aa1-c20849cc5669