IMDB Movie Analysis

PROJECT DESCRIPTION: -
The IMDb Movie Analysis project aims to explore and analyze a comprehensive dataset of movies available on the IMDb platform. This dataset contains essential information about movies, including director names, movie titles, duration, genre, budget, gross earnings, IMDb ratings, and more. Through in-depth data analysis using Excel, Data Visualization and Statistics techniques this project seeks to extract valuable insights and trends that contribute to a movie's success.

In this project, I was required to provide a detailed report for the below data record mentioning the answers of the questions that follows:

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.

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.

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.

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.

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.

MY APPROACH: -

I have reviewed the dataset and have a clear understanding of every column. After that, I noticed that there are 5043 rows and a total of 28 columns. This dataset has blank rows, null values, and unnecessary columns. Thus, I have made the decision to completely clean this dataset.

1) Firstly, I have removed the columns that don't relate to our project and don't offer any insightful information. Ultimately, I was left with just Ten columns: the name of the director, the length of the film, the genre, the budget, the gross, the IMDb rating, the language, title year and the country.
2) Then, I noticed that there were many blank rows. To find them I first clicked on “Find & Select” then clicked on “go to special” and selected the “blank” option. It highlighted all the blank rows. Then I clicked the Delete cell from home page and selected the “Entire rows delete” option. This process deleted the entire blank rows in the dataset.
3) Lastly, I also eliminated the dataset's duplicate rows. I now have 3836 Rows and 10 Columns overall. Here is the Cleaned Dataset for your use.

Cleaned Data set - Cleaned Data Set IMDB Analysis.csv

Hyperlink - Cleaned Data Set IMDB Analysis.csv

Drive link - https://drive.google.com/file/d/105-pIb1auLCRjv-XqKjL92OMkS2e2tM8/view?usp=drive_link

TECH STACK:

I ran the functions in Microsoft Excel 365 for this project in order to obtain the answers to the following questions. This was also how I plotted the graphs.

INSIGHTS:

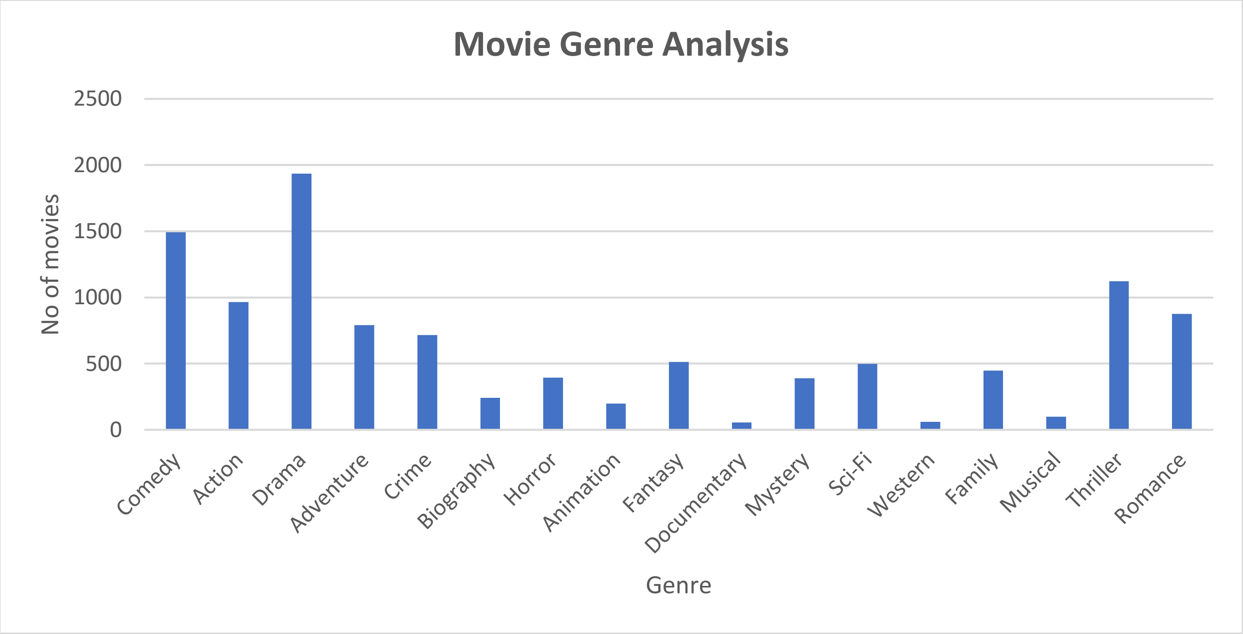
1.Movie Genre Analysis:

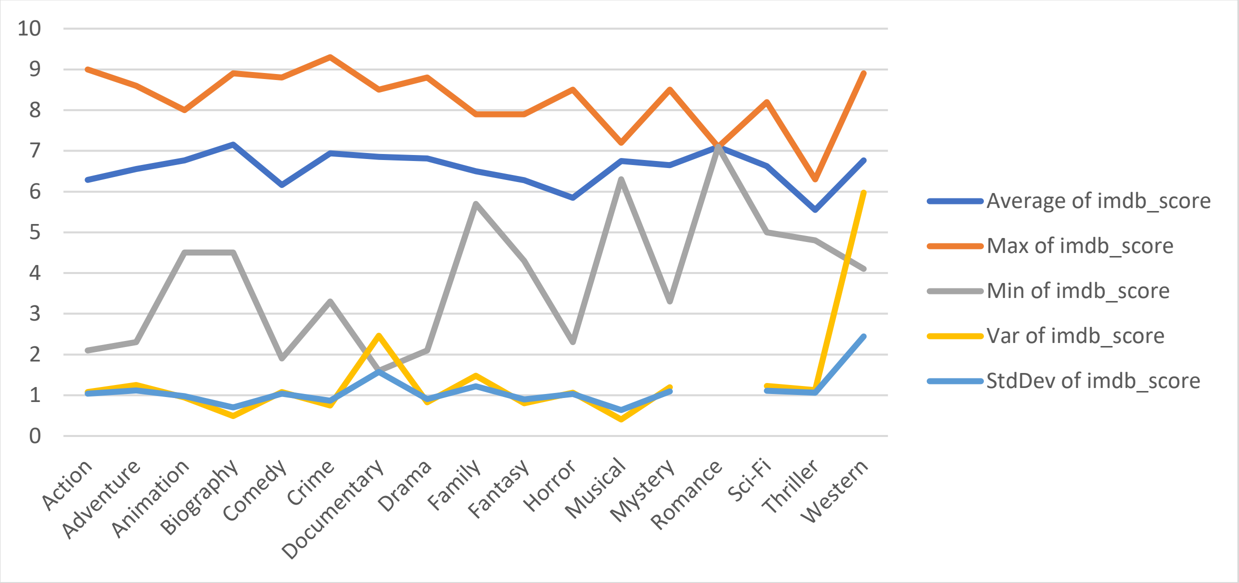
Task: Determine the most common genres of movies in the dataset. Then, for each genre, calculate descriptive statistics (mean, median, mode, range, variance, standard deviation of the IMDB scores.

Most Common Genre – Drama - Calculated using countif function
Mode – Mode(if())
Median – Median(if())
Average, max, Min, Var, Stdev – using pivot table

Genre	No of Movies	Median	Mode
Comedy	1492	6.3	6.4
Action	965	6.3	6.6
Drama	1935	6.9	6.7
Adventure	790	6.7	7.3
Crime	715	7	7.3
Biography	242	7.2	7
Horror	394	5.9	5.9
Animation	198	7	7.1
Fantasy	513	6.5	6.8
Documentary	54	7.4	7.5
Mystery	388	6.7	7.1
Sci-Fi	499	6.4	#N/A
Western	60	7.3	#N/A
Family	448	5.9	#N/A
Musical	98	6.75	#N/A
Thriller	1123	5.55	#N/A
Romance	875	7.1	#N/A

Main Genre	Average of imdb_score	Max of imdb_score	Min of imdb_score	Var of imdb_score	StdDev of imdb_score
Action	6.28746114	9	2.1	1.077778299	1.038161018
Adventure	6.552406417	8.6	2.3	1.251884274	1.118876344
Animation	6.763043478	8	4.5	0.945937198	0.972593028
Biography	7.153140097	8.9	4.5	0.487453684	0.698178834
Comedy	6.159921415	8.8	1.9	1.078549453	1.038532355
Crime	6.940077821	9.3	3.3	0.750223431	0.866154392
Documentary	6.855882353	8.5	1.6	2.463752228	1.569634425
Drama	6.812861272	8.8	2.1	0.824827112	0.90819993
Family	6.5	7.9	5.7	1.48	1.216552506
Fantasy	6.281081081	7.9	4.3	0.799354354	0.894066191
Horror	5.850909091	8.5	2.3	1.065197339	1.032083979
Musical	6.75	7.2	6.3	0.405	0.636396103
Mystery	6.652173913	8.5	3.3	1.193517787	1.092482396
Romance	7.1	7.1	7.1	#DIV/0!	#DIV/0!
Sci-Fi	6.628571429	8.2	5	1.225714286	1.107119815
Thriller	5.55	6.3	4.8	1.125	1.060660172
Western	6.766666667	8.9	4.1	5.973333333	2.444040371
Grand Total	**6.459984359**	**9.3**	**1.6**	**1.117754263**	**1.057238981**





Movie Genre Analysis.xlsx

Analysis Sheet:

Drive link:
https://docs.google.com/spreadsheets/d/1pbIfrOO8H6V0bWYl_EpjV2hMUXVHO_Qk/edit?usp=sharing&ouid=109690823837991827030&rtpof=true&sd=true

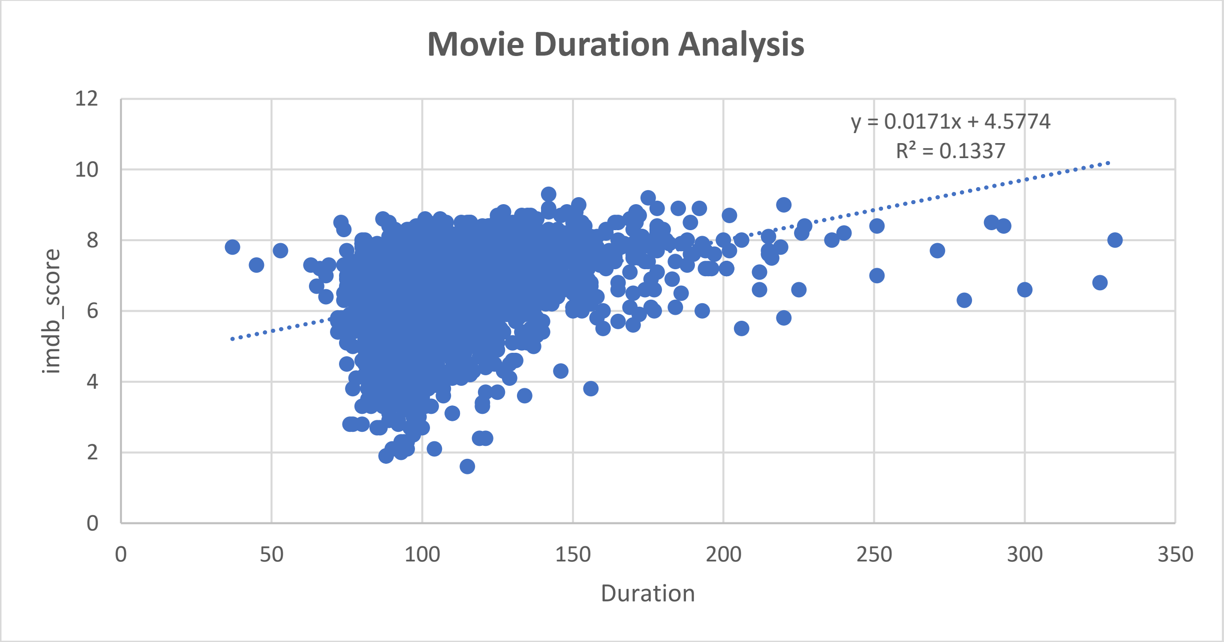
2.Movie Duration Analysis:

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

Calculated Mean, Median, Mode, Max, Min, variance, Standard deviation using excel function for Duration of movie.

Property	Value
Mean	110.0081
Median	106
Mode	101
Max	330
Min	37
variance	510.1702
standard deviation	22.58695

Scatter Graph plotted between Duration and IMDB Rating.



Drive Link:
https://docs.google.com/spreadsheets/d/1_w7dkwPeS_E6BmgYy-pyqBNzbj0uidMp/edit?usp=drive_link&ouid=109690823837991827030&rtpof=true&sd=true

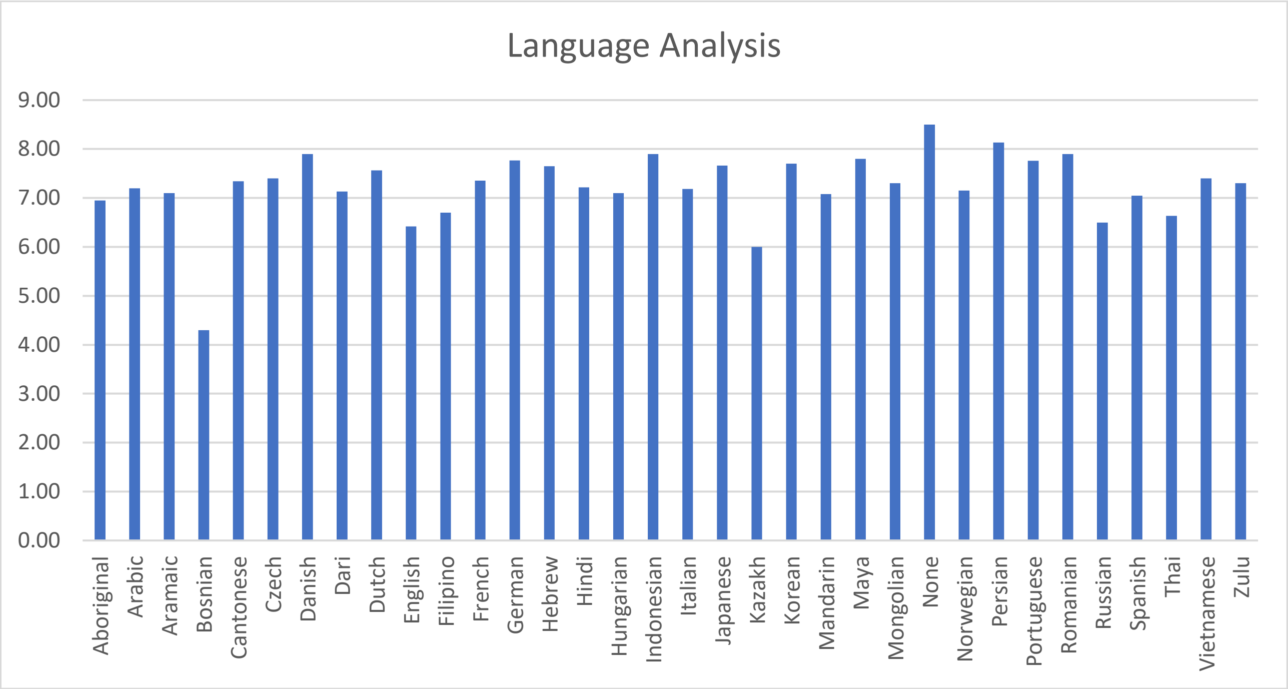
3.Movie Language Analysis:

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

Row Labels	Count of movie_title	Average of imdb_score	Max of imdb_score
Aboriginal	2	6.95	7.5
Arabic	1	7.2	7.2
Aramaic	1	7.1	7.1
Bosnian	1	4.3	4.3
Cantonese	7	7.342857143	7.8
Czech	1	7.4	7.4
Danish	3	7.9	8.3
Dari	2	7.5	7.6
Dutch	3	7.566666667	7.8
English	3675	6.421823129	9.3
Filipino	1	6.7	6.7
French	34	7.355882353	8.4
German	11	7.763636364	8.5
Hebrew	2	7.65	8
Hindi	5	7.22	8
Hungarian	1	7.1	7.1
Indonesian	2	7.9	8.2
Italian	7	7.185714286	8.9
Japanese	10	7.66	8.7
Kazakh	1	6	6
Korean	5	7.7	8.4
Mandarin	15	7.08	7.9
Maya	1	7.8	7.8
Mongolian	1	7.3	7.3
None	1	8.5	8.5
Norwegian	4	7.15	7.6
Persian	3	8.133333333	8.5
Portuguese	5	7.76	8.7
Romanian	1	7.9	7.9
Russian	1	6.5	6.5
Spanish	24	7.045833333	8.2
Thai	3	6.633333333	7.1
Vietnamese	1	7.4	7.4
Zulu	1	7.3	7.3

Min of imdb_score	Var of imdb_score	StdDev of imdb_score	Median
6.4	0.605	0.777817459	6.95
7.2	#DIV/0!	#DIV/0!	7.2
7.1	#DIV/0!	#DIV/0!	7.1
4.3	#DIV/0!	#DIV/0!	4.3
6.7	0.122857143	0.350509833	7.3
7.4	#DIV/0!	#DIV/0!	7.4
7.3	0.28	0.529150262	8.1
7.4	0.02	0.141421356	7.5
7.1	0.163333333	0.404145188	7.8
1.6	1.105930807	1.051632449	6.5
6.7	#DIV/0!	#DIV/0!	6.7
5.8	0.269812834	0.519435111	7.3
6.1	0.456545455	0.675681474	7.8
7.3	0.245	0.494974747	7.65
6	0.642	0.801249025	7.4
7.1	#DIV/0!	#DIV/0!	7.1
7.6	0.18	0.424264069	7.9
5.3	1.334761905	1.155318962	7
6	0.980444444	0.990173947	8
6	#DIV/0!	#DIV/0!	6
7	0.325	0.570087713	7.7
5.6	0.596	0.772010363	7.4
7.8	#DIV/0!	#DIV/0!	7.8
7.3	#DIV/0!	#DIV/0!	7.3
8.5	#DIV/0!	#DIV/0!	8.5
6.4	0.33	0.574456265	7.3
7.5	0.303333333	0.550757055	8.4
6.1	0.958	0.978774744	8
7.9	#DIV/0!	#DIV/0!	7.9
6.5	#DIV/0!	#DIV/0!	6.5
5.2	0.740851449	0.860727279	7.15
6.2	0.203333333	0.450924975	6.6
7.4	#DIV/0!	#DIV/0!	7.4
7.3	#DIV/0!	#DIV/0!	7.3

Graph Plotted between Languages and IMDB Rating.

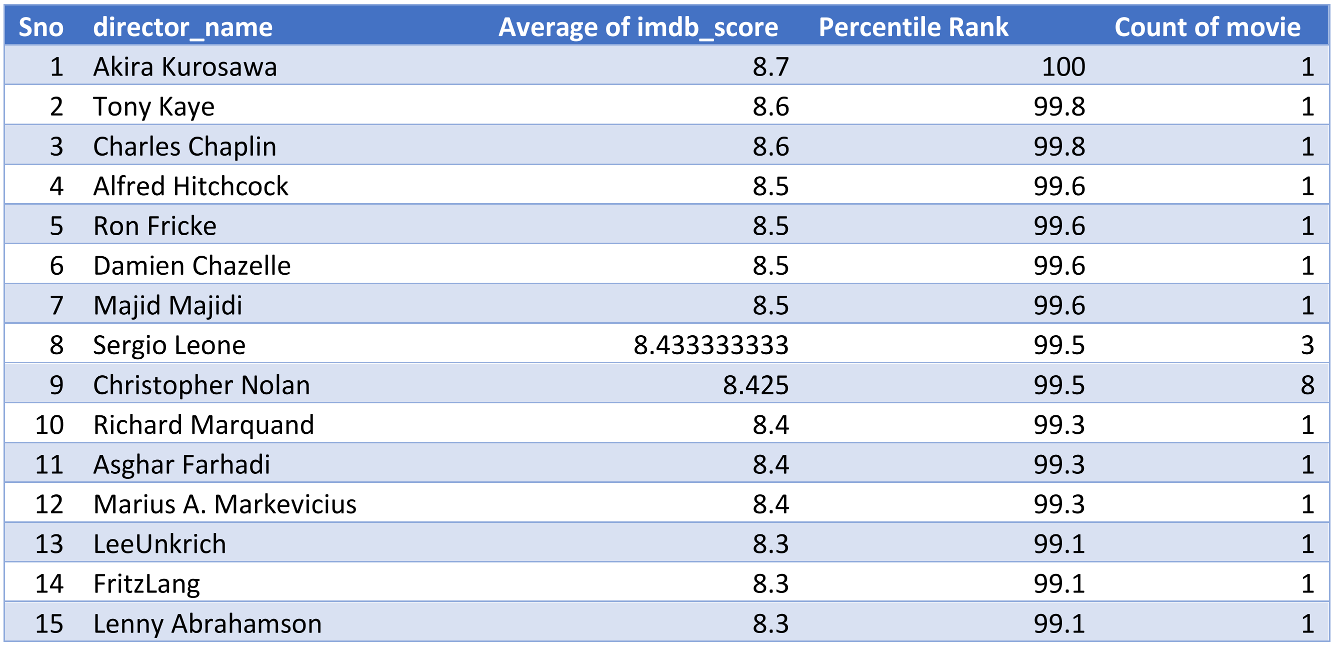


Analysis Sheet: Language Analysis.xlsx

Drive link:
https://docs.google.com/spreadsheets/d/1yeU3tbrhNpzcIkiAD59pM7exKlj-knrS/edit?usp=drive_link&ouid=109690823837991827030&rtpof=true&sd=true

4.Movie Director Analysis:

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



Sno	director_name	Average of imdb_score	Percentile Rank	Count of movie
1	Akira Kurosawa	8.7	100	1
2	Tony Kaye	8.6	99.8	1
3	Charles Chaplin	8.6	99.8	1
4	Alfred Hitchcock	8.5	99.6	1
5	Ron Fricke	8.5	99.6	1
6	Damien Chazelle	8.5	99.6	1
7	Majid Majidi	8.5	99.6	1
8	Sergio Leone	8.433333333	99.5	3
9	Christopher Nolan	8.425	99.5	8
10	Richard Marquand	8.4	99.3	1
11	Asghar Farhadi	8.4	99.3	1
12	Marius A. Markevicius	8.4	99.3	1
13	LeeUnkrich	8.3	99.1	1
14	FritzLang	8.3	99.1	1
15	Lenny Abrahamson	8.3	99.1	1

Director Analysis.xlsx

Analysis Sheet:

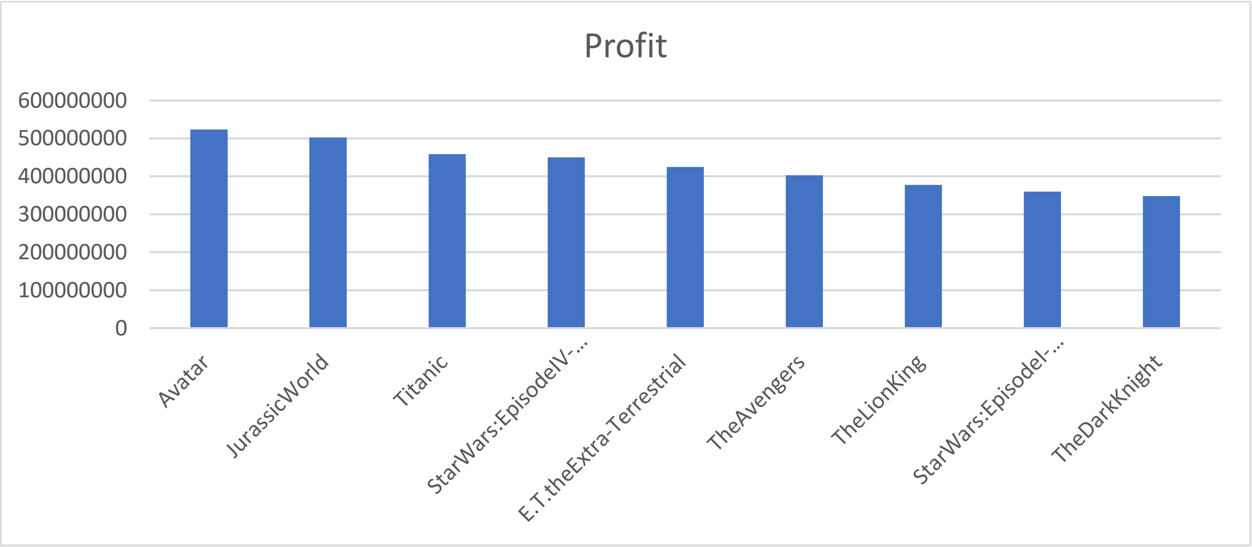
Drive Link:
https://docs.google.com/spreadsheets/d/13wriUxAGUz3iDcZxS8l0OxoFV2ApUNvY/edit?usp=drive_link&ouid=109690823837991827030&rtpof=true&sd=true

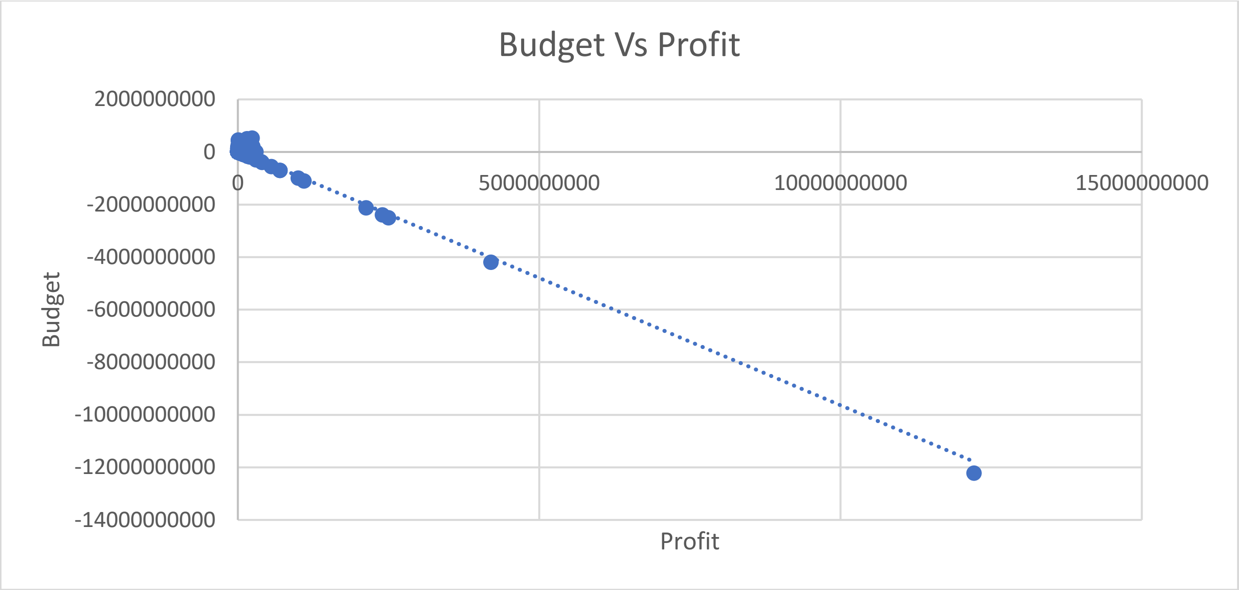
5.Movie Budget Analysis:

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

correlation coefficient	**0.1016651**

Sno	movie_title	Profit
1	Avatar	523505847
2	JurassicWorld	502177271
3	Titanic	458672302
4	StarWars:EpisodeIV-ANewHope	449935665
5	E.T.theExtra-Terrestrial	424449459
6	TheAvengers	403279547
7	TheLionKing	377783777
8	StarWars:EpisodeI-ThePhantomMenace	359544677
9	TheDarkKnight	348316061
10	TheHungerGames	329999255





Budget Analysis.xlsx

Analysis Sheet:

Drive link:
https://docs.google.com/spreadsheets/d/1uX6ksNKudk_tEKV_DBVStbU6UQPtK3JQ/edit?usp=drive_link&ouid=109690823837991827030&rtpof=true&sd=true

The Results Dataset Link:

https://drive.google.com/drive/folders/19V9jjKufOJB116PiE6Na4tDFq10kD7qu?usp=sharing

I have noticed that,

- The Most common movie genres from the dataset are Drama, Comedy, Thriller and Action.
- The Average duration of a Movie is 109 minutes. The trendline between the duration vs imdb score is elevated upward with R^2 = 0.1337
- The Most common languages used in the movies are English, French, Spanish, Mandarin and German. I have also Observed that the languages Telugu and Persian have the highest average imdb score.
- I have identified that Akira Kurosawa, Tony Kaye, Charles Chaplin, Alfred Hitchcock, Ron Fricke, Damien Chazelle, Majid Majidi, Sergio Leone, Christopher Nolan,
- Richard Marquand, Asghar Farhadi, Marius A. Markovic, Lee Unkrich, Fritzl Ang, Lenny Abrahamson. are the top 15 directors with average imdb score >=8.3
- The Top-5 with highest profits is Avatar, Jurassic World, Titanic, Star Wars: Episode IV - A New Hope and E.T. The Extra-Terrestrial. The Correlation between budget and gross is positive.

RESULTS:

Thanks to this project, I have gained a lot of knowledge about data analysis using Excel's data visualization capabilities and statistical know-how. I've learned how to apply my data analysis expertise to address real-world problems thanks to this.

All Data sets Drive Link:
https://drive.google.com/drive/folders/1Hg0HkG6NKWhOL6UeCnIKRsD8Nxt_kcx_?usp=sharing