IMDB Movie Analysis

Link:- https://docs.google.com/spreadsheets/d/14XdMr7v7Af-KMCuqsOqTP2UwMpQoeF99/edit#gid=1567918698

❖ Project Description:

The main goal of this project is to analyze a movies dataset and identify the factors that contribute to a movie's success. This analysis aims to uncover how various factors such as genre, duration, budget, etc., influence IMDb ratings. The insights gained from this analysis will assist investors, producers, and directors in making data-driven decisions.

❖ Approach:

- Download the dataset.
- Gain a comprehensive understanding of the data.
- Clean the data by removing null and duplicate columns, deleting unnecessary columns, and eliminating special characters.
- Utilize Excel and statistical formulas to address the project's objectives.
- Create visualizations to extract meaningful insights from the data.

❖ Tech Stack Used:

- Microsoft Excel (2019): Used for data analysis and visualization.
- Microsoft Word: Used to create a presentation report summarizing the findings of the analysis.

❖ Data Cleansing:

- Deleting unwanted rows, colour, director_facebook_likes, actor_3_facebook_likes, actor_2_name, actor_1_facebook_likes, facenumber_in_poster, movie_imdb_link, content_rating, actor_2_facebook_likes, cast_total_facebook_likes, actor_3_name.

- Deleted rows which had incomplete information from columns Director's Name, Gross ,Budget, Plot_keywords, rather than deleting rows having no aspect ratio, all the blank cells in aspect ratio were replaced by most common ratio of 2.35 and 1.85 which was found by unique and countif function.

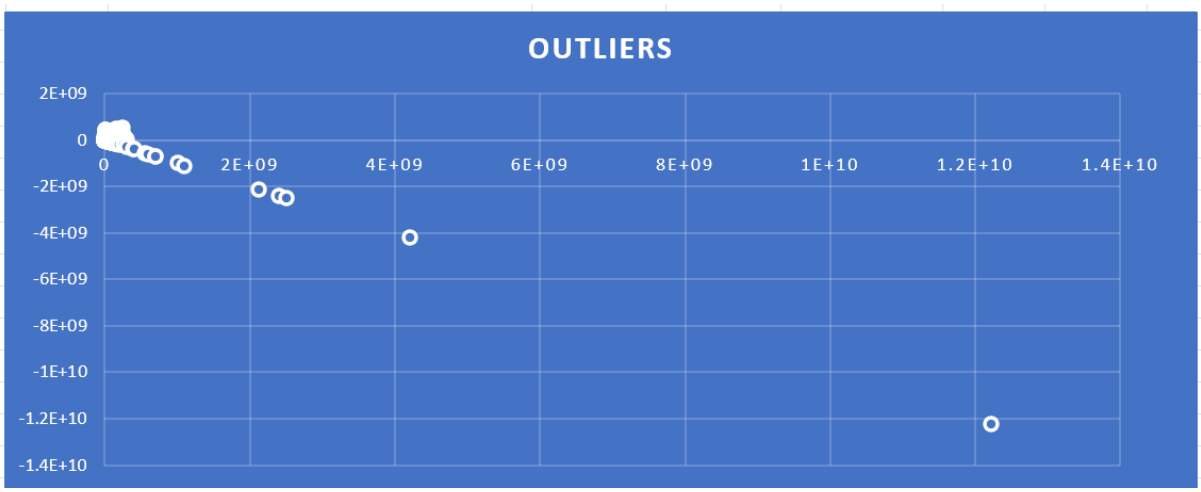
- 35 Duplicates were removed 3825 rows remains.

- Excel sheet is sorted using sort function in Data tab making the sheet in descending order of Profit per movie.

- Languages for 3 movies were missing. Because the country is USA, the blank spaces were replaced by English Language.

❖ Outlier Found

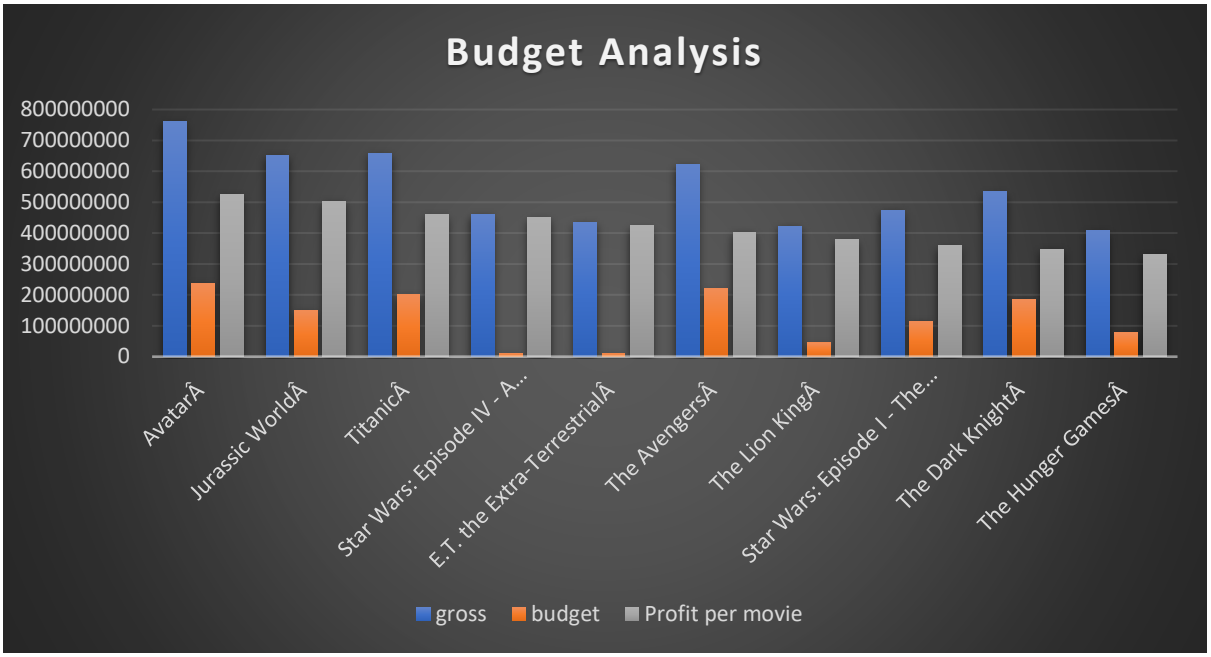
	-
First Quartile Q1	10241539
Second Quartile Q1(Median)	1062898
Third Quartile Q1	24833860
IQR	35075399



Outliers
-2127109510
-2397701809
-2499804112
-4199788333
-12213298588

Top movies

Movie_Title	gross	budget	Profit per movie
AvatarÂ	760505847	237000000	523505847
Jurassic WorldÂ	652177271	150000000	502177271
TitanicÂ	658672302	200000000	458672302
Star Wars: Episode IV - A New HopeÂ	460935665	11000000	449935665
E.T. the Extra-TerrestrialÂ	434949459	10500000	424449459
The AvengersÂ	623279547	220000000	403279547
The Lion KingÂ	422783777	45000000	377783777
Star Wars: Episode I - The Phantom MenaceÂ	474544677	115000000	359544677
The Dark KnightÂ	533316061	185000000	348316061
The Hunger GamesÂ	407999255	78000000	329999255



Findings:-

Correlation Coefficient for Top 10 movies is 0.86

A correlation coefficient of 0.86 indicates a strong positive linear relationship between Gross and Budget

Correlation Coefficient for all the Movies is 0.100

This depicts weak linear relationship between Gross and Budget.

Avatar movie has the highest profit. Star wars and Extra Terrestrial had lowest Budget but made huge profits

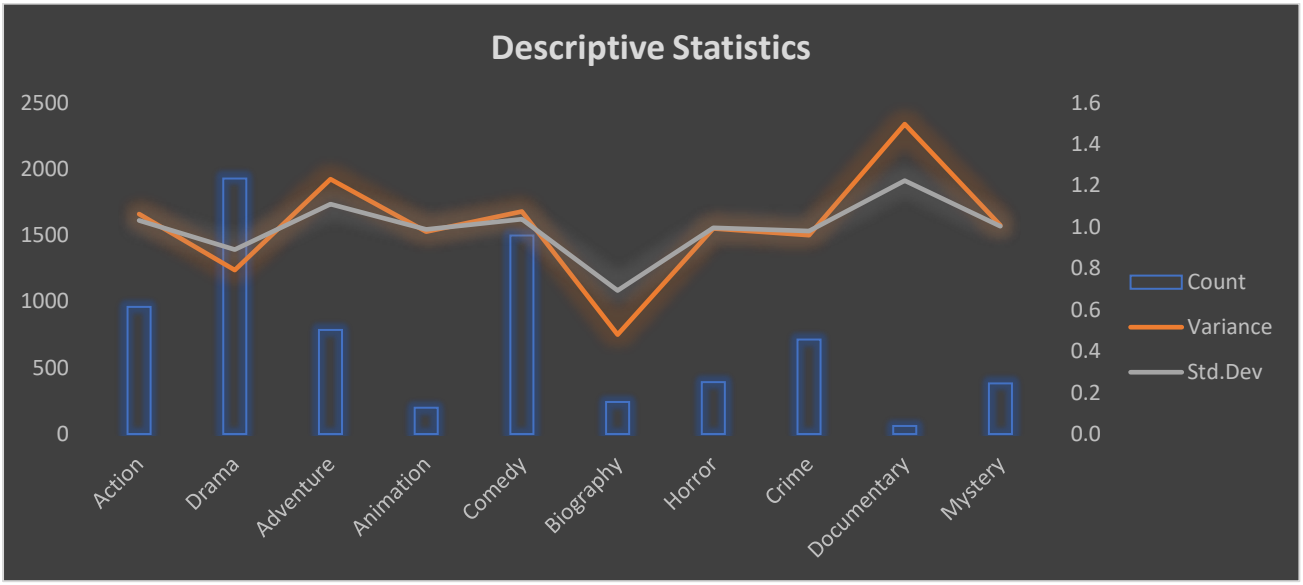
❖ Genre Analysis:-

Analyzing 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

Descriptive Statistics

Genre	Count	Average of imbd_score	Profitability	Variance	Std.Dev	Median	Range
Action	958	6.3	5013129.7	1.1	1.0	6.35	6.9
Drama	1928	6.8	-3291152.6	0.8	0.9	6.9	7.2
Adventure	784	6.5	13364413.7	1.2	1.1	6.6	6.6
Animation	198	6.7	1115491.5	1.0	1.0	6.8	5.8
Comedy	1497	6.2	8214214.8	1.1	1.0	6.3	6.9
Biography	241	7.2	7831660.4	0.5	0.7	7.2	4.4
Horror	391	5.9	-18042006.7	1.0	1.0	6	6.3
Crime	712	6.5	705359.1	1.0	1.0	6.6	6.9
Documentary	61	7.0	7576628.2	1.5	1.2	7.3	6.9
Mystery	382	6.5	10599587.4	1.0	1.0	6.5	5.5



Findings:-

Drama is one of the most popular genre that has appeared 1928 movies with an average imdb rating of 6.8 followed by Comedy with an average imdb rating of 6.2.

❖ Movie Duration Analysis:-

Analyze the distribution of movie durations and its impact on the IMDB score.

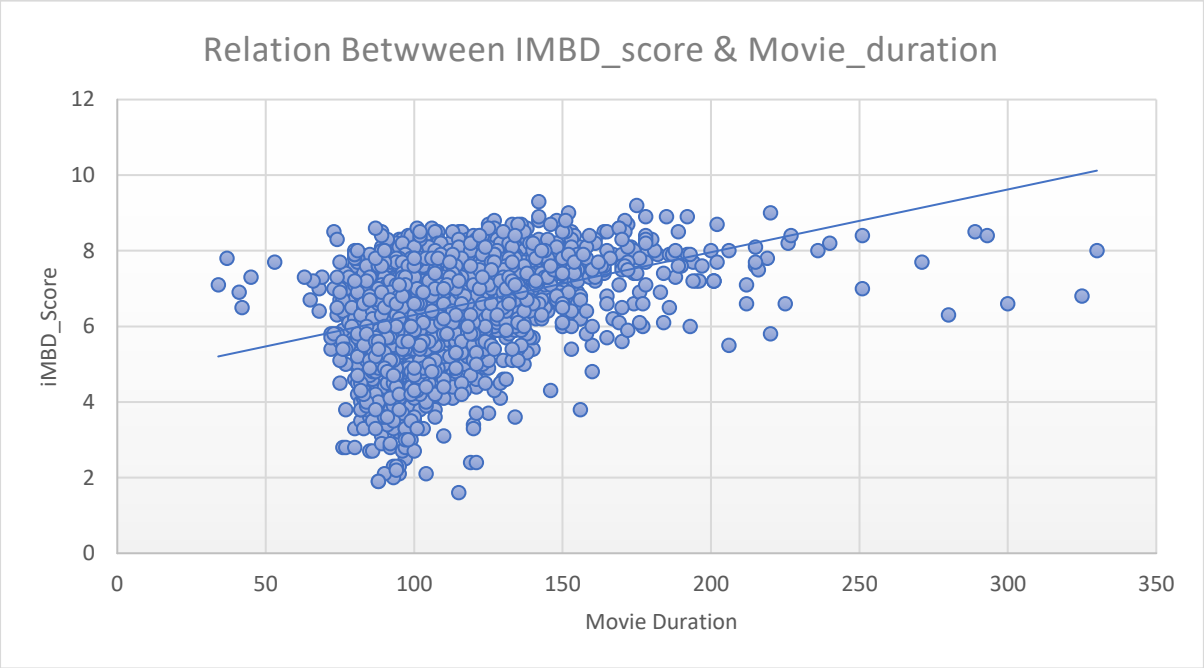
Task: Analyze the distribution of movie durations and identify the relationship between movie duration and IMDB score.

Descriptive Statistics

Movie Duration	Values
Avg Duration	109.99
Median	106.00
Std.Dev	22.77
Mode	101.00
Variance	518.57
Short Movie Duration	34
Long Movie Duration	330

Findings:-

Correlation Between Imbd_Score and Movie Duration close to 0 indicate a weak or no linear relationship.



Findings:-

Duration of 80 to 130 has got the maximum films and the imdb score lies between 4.5 –8.5

Shortest movie duration is of 34 min whereas longest movie duration is of 330 min.

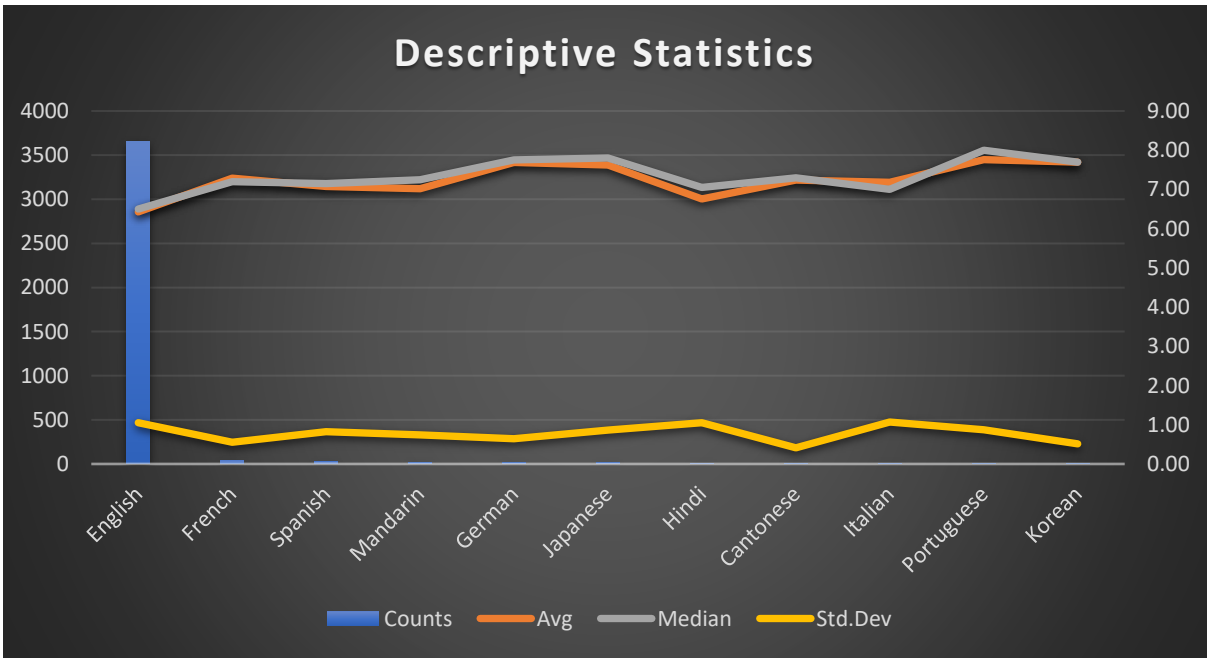
❖ Language Analysis:

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.

Sorted in Descending Order

Language	Counts	Avg	Median	Std.Dev
English	3649	6.43	6.5	1.05
French	37	7.29	7.2	0.55
Spanish	24	7.08	7.15	0.82
Mandarin	14	7.02	7.25	0.74
German	12	7.69	7.75	0.64
Japanese	12	7.63	7.8	0.86
Hindi	10	6.76	7.05	1.05
Cantonese	8	7.24	7.3	0.41
Italian	7	7.19	7	1.07
Portuguese	5	7.76	8	0.88
Korean	5	7.70	7.7	0.51



Findings:-

English is the language used in maximum of movies. The avg imbd_score is 6.43.

Persian and Telugu are languages with more than 8 imbd_score.

From the top 10 movies French, Spanish, mandarin, German, Japanese, Cantonese, Italian, Portuguese, Korean have higher imbd_score than English.

This is due to consistent audience because of fewer movies in these languages.

The Average of movie ratings are consistent across languages ranging from 6.4 –7.7

❖ 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.

Top Directors with highest average imdb ratings

Rank	Ranking Directors as per their IMBD_Score		90th Percentile
1	Tony Kaye	8.60	7.5
1	Charles Chaplin	8.60	7.5
3	Alfred Hitchcock	8.50	7.5
3	Damien Chazelle	8.50	7.5
3	Majid Majidi	8.50	7.5
3	Ron Fricke	8.50	7.5
7	Sergio Leone	8.43	7.5
8	Christopher Nolan	8.43	7.5
9	Richard Marquand	8.40	7.5
9	Asghar Farhadi	8.40	7.5

Top Directors with most number of movies

Directors	Count of Movies	Avg IMBD_Score
Steven Spielberg	25	7.54
Clint Eastwood	19	7.21
Woody Allen	19	7.00
Ridley Scott	17	7.07
Tim Burton	16	6.93
Steven Soderbergh	16	6.71
Martin Scorsese	16	7.68
Renny Harlin	15	5.75
Spike Lee	15	6.73

Findings:-

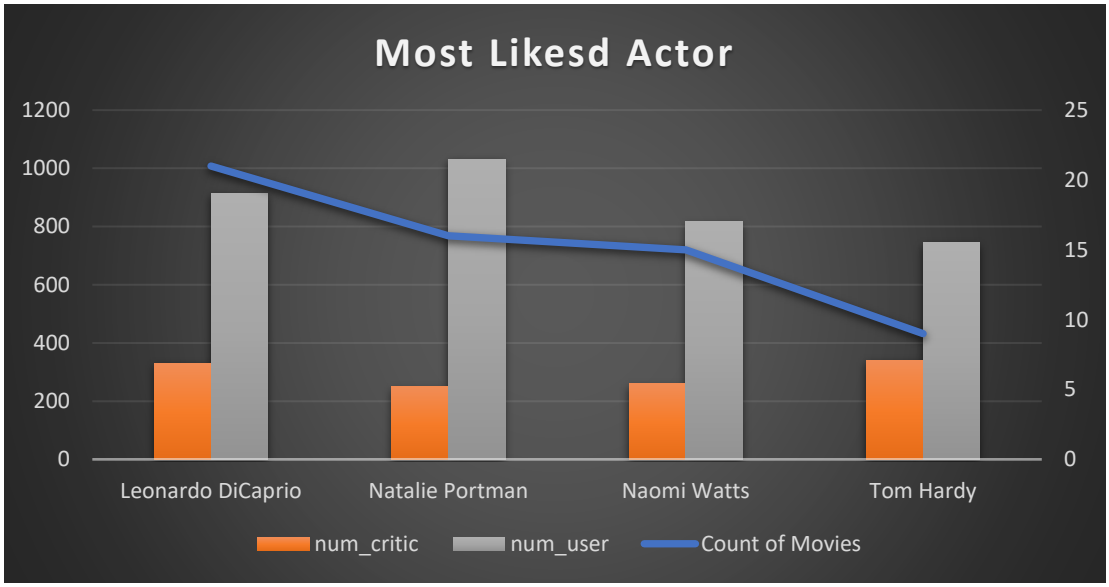
Charles Chaplin and Tony Kaye have the highest average IMDb score of 8.60, with only 1 movie.
Steven Spielberg has the highest average imdb ratings of 7.54 for a total of 25 movies indicating a consistent record.
Percentile : Each director's average IMDb score is compared against a common benchmark of 90th percentile, to know their relative position in dataset.

❖ Actor_1_Analysis:-

Following are the actors mostly appearing in the Movies

Actor_1	Count of Movies
Robert De Niro	42
Johnny Depp	38
J.K. Simmons	31
Nicolas Cage	30
Denzel Washington	30
Bruce Willis	29
Matt Damon	29
Robert Downey Jr.	26
Robin Williams	26
Liam Neeson	26

Actor_1	Count of Movies	num_critic	num_user
Leonardo DiCaprio	21	330.19	914.48
Natalie Portman	16	249.13	1031.94
Naomi Watts	15	259.40	816.80
Tom Hardy	9	341.33	744.11



Findings:-

From the Actor_1 Analysis we can understand that **Leonardo DiCaprio** is Critics and Users Favourite Actor.

❖ Conclusion:-

In this IMDb Movie Analysis project, I've developed various logical, statistical, and technical skills to derive meaningful insights from the dataset. Concepts such as calculating averages, creating frequency tables, and identifying outliers have enabled me to deepen my understanding of the data and enhance my ability to analyze it effectively.

By applying statistical methods and leveraging the technical capabilities of Microsoft Excel, I've been able to streamline data analysis tasks and simplify complex calculations. Excel's features have expedited the analytics process, allowing for more efficient data manipulation and interpretation.

Furthermore, I've learned the importance of visualizing data through graphs and charts, which significantly improves data comprehension. Visual representations enhance the clarity and impact of insights derived from the data, making it easier to communicate findings and understand trends.

Through this project, I've gained valuable knowledge on selecting appropriate visualization techniques based on the dataset and desired outcomes, enhancing my ability to present and interpret data effectively. This experience has equipped me with essential skills for conducting data-driven analyses and drawing actionable conclusions from datasets.

_____**********_____

_____**********_____