IMDB MOVIE ANALYSIS

Project Description:

Movie Database (IMDB) provides information on a variety of movie-related topics, including box office results, popular genres, the influence of actors and directors, and critical acclaim.

Data on movie titles, release dates, box office receipts, ratings, and cast and crew members are among the details that need to be collected and cleaned for the project.

The analysis findings can offer producers, distributors, and movie studios useful information about how to make popular films that appeal to audiences. Movie buffs can also utilize the data to research their preferred films, actresses, and directors and develop a better grasp of the film business.

The project can be enhanced further to incorporate user review sentiment analysis, box office success prediction, and analysis of regional variances in movie preferences. Overall, the IMDB Movie Analysis project offers a thorough method for studying the film industry and can deliver insightful information to a variety of stakeholders.

The dataset in question has 28 columns and 5044 rows.

I've cleaned up the data and eliminated duplicate and null values in accordance with the provided question.

I looked at the data set and came to my conclusions.

I have used the Five 'Whys' method of root cause analysis.

By making new columns, pivot tables, and graphs, I have answered the questions.

Approach:

After downloading the XLS file of IMDB_MOVIES, I cleaned the Data Set by removing bank rows, and null values and dropping some columns. Then, I analyzed the data and gave the answer to the question by using filters, Pivot Tables, Graphs, and different formulas.

At last, I created this report in MS Word and then convert it into a pdf file to submit the report.

Tech-Stack Used:

MS Excel for analysis and MS Word to create report.

Insights:

I understand Box office performance, Audience preferences, Critical reception, Actor and director influence from the IMDB Data Set.

A. **Cleaning the data:** It is one of the most important steps to perform before doing the analysis. Using knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Process for Cleaning the data:

Initially, there are 5044 rows and 28 columns.

I have deleted the blank rows using the following process:

Ctrl+G ->Select Blanks -> Select Delete Cell -> Delete Sheet Rows

I have selected each column, applied sort and filter, then deleted null values. After doing these operations there are 3757 rows left.

I have dropped the following column which are not relevant for analysis.

a) Colour b) Movie_ imdb_link c) Aspect Ratio d) Duration.

So, after cleaning the data we have 3757 rows and 24 columns.

Cleaned data file link:

https://docs.google.com/spreadsheets/d/13OBZm7s2-006G0ZSvbkbXjGvPQNZK5qm/edit?usp=sharing&ouid=104660294464707081800&rtpof=true&sd=true

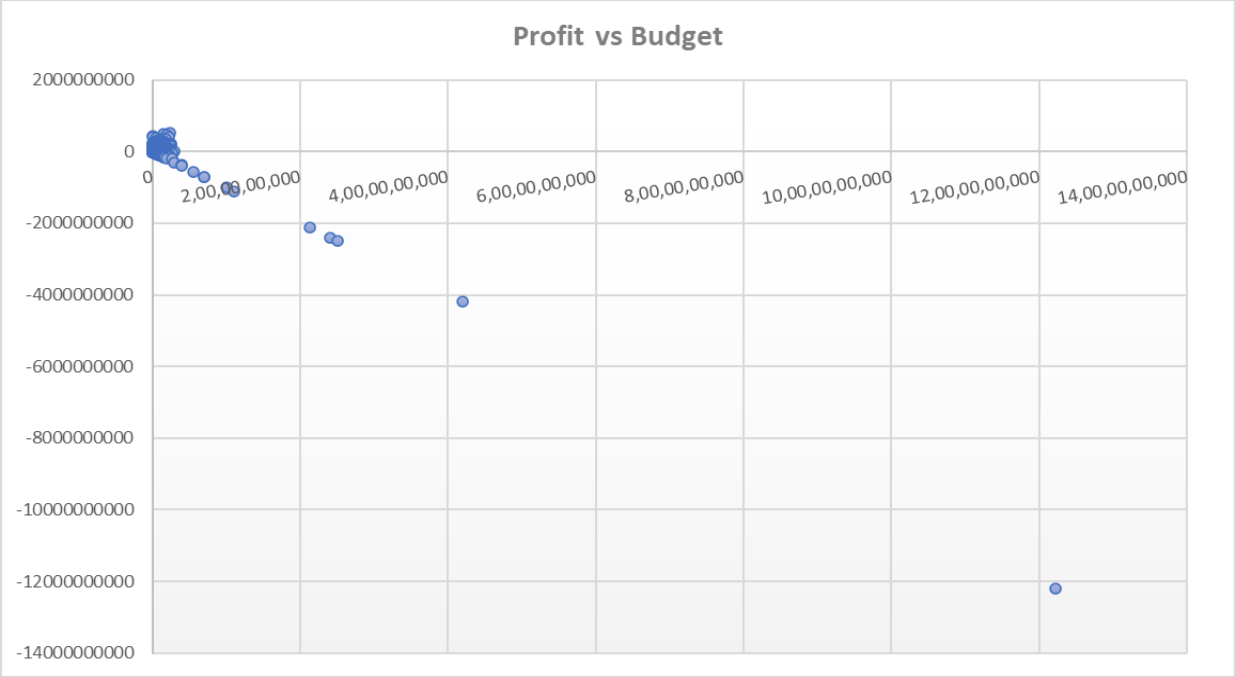
Working file link:

https://docs.google.com/spreadsheets/d/13OBZm7s2-006G0ZSvbkbXjGvPQNZK5qm/edit?usp=sharing&ouid=104660294464707081800&rtpof=true&sd=true

(Open in MS Excel)

B. **Movies with highest profit:** Create a new column called profit which contains the difference between the two columns: gross and budget. Sort the column using the profit column as a reference. Plot profit (y-axis) vs budget (x-axis) and observe the outliers using the appropriate chart type.
Your task: Find the movies with the highest profit.

Scatterplot:



TOP 5 PROFITABLE MOVIES:

director_name	actor_1_name	movie_title	title_year	imdb_score	Profit
James Cameron	CCH Pounder	Avatar	2009	7.9	523505847
Colin Trevorrow	Bryce Dallas Howard	Jurassic World	2015	7	502177271
James Cameron	Leonardo DiCaprio	Titanic	1997	7.7	458672302
George Lucas	Harrison Ford	Star Wars: Episode IV – A New Hope	1977	8.7	449935665
Steven Spielberg	Henry Thomas	E.T. the Extra-Terrestrial	1982	7.9	424449459

Avatar is the movie with the **highest** profit of **523505847** & Its budget was **237000000** and its gross income was **760505847**.

C. **Top 250:** Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film.
Your task: Find IMDB Top 250

Rank	movie_title	num_voted_us	imdb_score
1	The Shawshank RedemptionÂ	1689764	9.3
2	The Dark KnightÂ	1676169	9
3	InceptionÂ	1468200	8.8
4	Fight ClubÂ	1347461	8.8
5	Pulp FictionÂ	1324680	8.9
6	Forrest GumpÂ	1251222	8.8
7	The Lord of the Rings: The Fellowship of the RingÂ	1238746	8.8
8	The MatrixÂ	1217752	8.7
9	The Lord of the Rings: The Return of the KingÂ	1215718	8.9
10	The GodfatherÂ	1155770	9.2
11	The Dark Knight RisesÂ	1144337	8.5
12	The Lord of the Rings: The Two TowersÂ	1100446	8.7
13	Se7enÂ	1023511	8.6
14	The AvengersÂ	995415	8.1
14	The AvengersÂ	995415	8.1
16	GladiatorÂ	982637	8.5
17	Batman BeginsÂ	980946	8.3
18	Django UnchainedÂ	955174	8.5
19	InterstellarÂ	928227	8.6
20	Star Wars: Episode IV - A New HopeÂ	911097	8.7
21	The Silence of the LambsÂ	887467	8.6
22	AvatarÂ	886204	7.9
23	Inglourious BasterdsÂ	885175	8.3
24	Saving Private RyanÂ	881236	8.6
25	The DepartedÂ	873649	8.5
26	Schindler's ListÂ	865020	8.9
27	MementoÂ	845580	8.5

240	The Pursuit of HappynessÂ	338383	8	11036
241	ElysiumÂ	338087	6.6	17689
242	Live Free or Die HardÂ	336235	7.2	13961
243	War of the WorldsÂ	334345	6.5	12758
244	Indiana Jones and the Kingdom of the Crystal SkullÂ	333847	6.2	14959
245	CasinoÂ	333542	8.2	24183
246	The Day After TomorrowÂ	333248	6.4	20553
247	WarriorÂ	332276	8.2	29692
248	The Hurt LockerÂ	332065	7.6	11114
249	Quantum of SolaceÂ	330784	6.7	2023
250	The Girl with the Dragon TattooÂ	330152	7.8	20388
251	Indiana Jones and the Temple of DoomÂ	329969	7.6	11898
252	Ice AgeÂ	328159	7.6	5437
253	The WolverineÂ	328067	6.7	23755
254	LucyÂ	327367	6.4	32325
255	Dallas Buyers ClubÂ	326494	8	17738
256	The Incredible HulkÂ	326286	6.8	5811
257	Transformers: Dark of the MoonÂ	326180	6.3	2593
258	American SniperÂ	325264	7.3	16277
259	American GangsterÂ	324671	7.8	20354
260	Captain PhillipsÂ	323353	7.9	16281

IMDb_Top_250 column which are not in the English.

Rank	movie_title	num_voted_us	imdb_score	language
1	AmÃ©lieÂ	534262	8.4	French
2	City of GodÂ	533200	8.7	Portuguese
3	The Good, the Bad and the UglyÂ	503509	8.9	Italian
4	Pan's LabyrinthÂ	467234	8.2	Spanish
5	Spirited AwayÂ	417971	8.6	Japanese
6	OldboyÂ	356181	8.4	Korean
7	The Lives of OthersÂ	259379	8.5	German
8	DownfallÂ	248354	8.3	German
9	ApocalyptoÂ	236000	7.8	Maya
10	Seven SamuraiÂ	229012	8.7	Japanese
11	Princess MononokeÂ	221552	8.4	Japanese
12	Crouching Tiger, Hidden DragonÂ	217740	7.9	Mandarin
13	Howl's Moving CastleÂ	214091	8.2	Japanese
14	The Passion of the ChristÂ	179235	7.1	Aramaic
15	Amores PerrosÂ	173551	8.1	Spanish
16	The HuntÂ	170155	8.3	Danish
17	Das BootÂ	168203	8.4	German
18	Run Lola RunÂ	161471	7.8	German
19	A SeparationÂ	151812	8.4	Persian
20	HeroÂ	149414	7.9	Mandarin
20	HeroÂ	149414	7.9	Mandarin
22	The Raid: RedemptionÂ	148221	7.6	Indonesian
23	A Fistful of DollarsÂ	147566	8	Italian
24	Letters from Iwo JimaÂ	132149	7.9	Japanese
25	The Secret in Their EyesÂ	131831	8.2	Spanish
26	The OrphanageÂ	120189	7.5	Spanish
27	Good Bye Lenin!Â	114407	7.7	German

132	The Names of LoveÂ	6304	7.2	French
133	I Served the King of EnglandÂ	6183	7.4	Czech
134	AddictedÂ	5975	5.2	Spanish
135	Sleep DealerÂ	5699	5.9	Spanish
136	FatelessÂ	5603	7.1	Hungarian
137	Godzilla 2000Â	5442	6	Japanese
138	MoliÃ¨reÂ	5166	7.3	French
139	The Widow of Saint-PierreÂ	4767	7.3	French
140	1911Â	4670	6	Mandarin
141	The CircleÂ	4555	7.5	Persian
142	Mississippi MermaidÂ	4391	7.2	French
143	Bon voyageÂ	4293	6.9	French
144	The Adventures of PinocchioÂ	4086	5.3	Italian
145	CleanÂ	3924	6.9	French
146	QuinceaÃ±eraÂ	3675	7.1	Spanish
147	For Greater Glory: The True Story of CristiadaÂ	3665	6.6	Spanish
148	Remember Me, My LoveÂ	3548	6.5	Italian
149	Nomad: The WarriorÂ	3322	6	Kazakh
150	When the Cat's AwayÂ	2843	6.9	French
151	The Holy GirlÂ	2720	6.7	Spanish
152	TangoÂ	2412	7.2	Spanish
153	Futuro BeachÂ	1738	6.1	Portuguese
154	The Legend of SuriyothaiÂ	1666	6.6	Thai
155	R100Â	1658	6.1	Japanese
156	La otra conquistaÂ	1024	6.8	Spanish
157	One to AnotherÂ	1010	5.8	French
158	Journey from the FallÂ	775	7.4	Vietnamese

D. **Best Directors:** TGroup the column using the director_name column.

Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10 director. In case of a tie in IMDb score between two directors, sort them alphabetically.
Your task: Find the best directors.

Row Labels	Average of imdb_score
Akira Kurosawa	8.7
Tony Kaye	8.6
Charles Chaplin	8.6
Alfred Hitchcock	8.5
Ron Fricke	8.5
Majid Majidi	8.5
Damien Chazelle	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
Asghar Farhadi	8.4
Grand Total	**8.47**

E. **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.
Your task: Find popular genres.

Genres	Count of Genres
Comedy|Drama|Romance	147
Drama	141
Comedy	138
Comedy|Drama	138
Comedy|Romance	131
Drama|Romance	115
Crime|Drama|Thriller	82
Action|Crime|Thriller	56
Action|Crime|Drama|Thriller	50
Action|Adventure|Sci-Fi	48

We can see that Comedy I Drama I Romance is the most popular genre.

F. Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

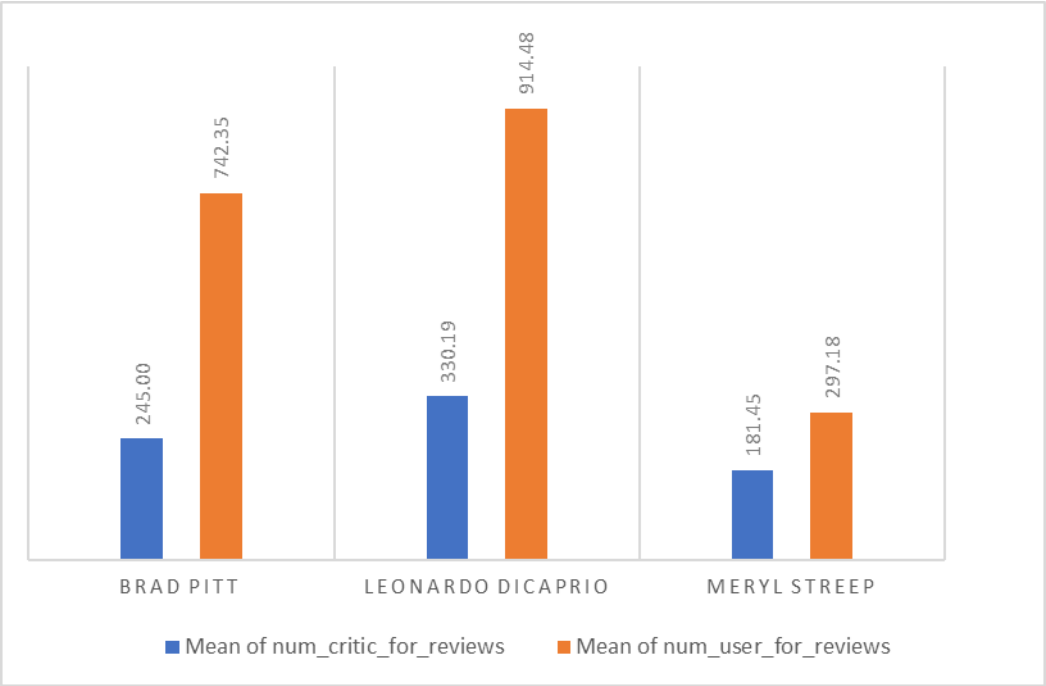
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the critic-favorite and audience-favorite actors.

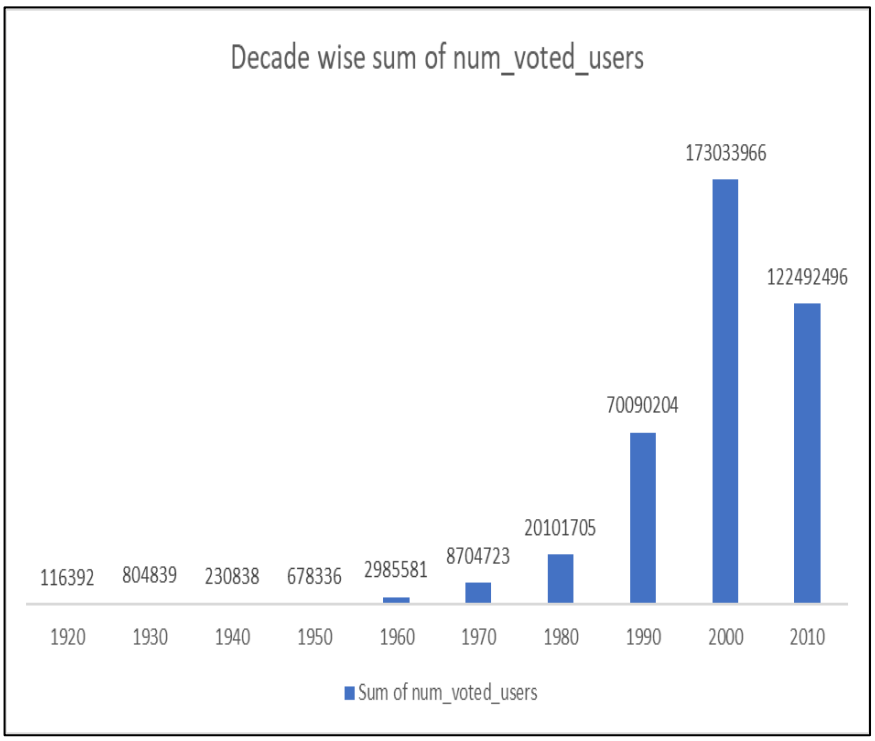
Row Labels	Mean of num_critic_for_reviews	Mean of num_user_for_reviews
Brad Pitt	245.00	742.35
Leonardo DiCaprio	330.19	914.48
Meryl Streep	181.45	297.18



Here we can see that Leonardo DiCaprio is having the highest mean of critics and reviews of the audience.

User Voting by decade by using Pivot Table.

Decade	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336
1960	2985581
1970	8704723
1980	20101705
1990	70090204
2000	173033966
2010	122492496



Result:

In this project of IMDB Movie Analysis, I had applied statistics and MS-Excel's technical abilities to analyze the given data. MS Excel simplifies and makes data into a structured format which makes it easy to understand. Using filters, sorting, charts, pivot tables, and graphs helped me to answer the questions as it was easy to visualize the data.

I have observed the following:

1. The relationship between movie ratings and box office earnings.

2. A comparison of ratings for various directors, actors, or genres.

3. Analysis of movie rating trends.

4. A breakdown of the top films by geography, language, or other demographics.

5. Determining the most significant directors or performers based on their contribution to highly rated films.