IMDB Movie Analysis

Project Description:

Introduction:

In this project, we analyzed a dataset containing information about various movies from IMDB. We aimed to explore this dataset and derive insights by framing the problem, cleaning the data, and using data analysis techniques.

Problem Framing:

Our goal was to frame a problem by asking some key questions such as What do you see happening? What is the impact of the problem on stakeholders? What is the impact of the problem not being solved? After brainstorming, we came up with the following problem: What factors contribute to a movie's success? We defined success based on the movie's profit and rating.

Approach:

The project will begin with cleaning the data, dropping unnecessary columns, and removing null values. Next, the analysis will focus on finding movies with the highest profit by creating a new column called "profit" and sorting the dataset by this column. Outliers will be identified using appropriate chart types.

The project will also determine the top 250 movies based on the IMDb rating and number of voted users. A new column called "IMDb_Top_250" will be created to store these movies, and a rank column will be added to indicate the ranks of the corresponding films. The analysis will also extract all non-English movies from the top 250 list and store them in a new column named "Top_Foreign_Lang_Film".

The next step of the project will focus on finding the best directors based on the mean IMDb score of their movies. The analysis will group the dataset using the director_name column and identify the top 10 directors with the highest mean IMDb scores. In case of a tie in IMDb score between two directors, the analysis will sort them alphabetically.

The project will also determine the popular genres in the movie industry by grouping the data based on genre and finding the mean num_voted_users of each genre. Three new columns will be created to identify movies in which "Meryl Streep", "Leonardo DiCaprio", and "Brad Pitt" are the lead actors, and the rows of all these columns will be appended to create a new column named "Combined". The analysis will group the combined column using the actor_1_name column and find the actors with the highest mean of num_critic_for_reviews and num_users_for_review.

Finally, the project will create a bar chart to observe the change in the number of voted users over the decades. A new column called "decade" will be created to represent the decade to which each movie belongs, and the column will be sorted based on this information. The dataset will be grouped by decade, and the sum of users voted in each decade will be stored in a new data frame called "df_by_decade".

Tech-Stack Used:

The analysis was performed using Microsoft Excel 2021. Pivot Table are used to filter and sort the data. Excel charts and graphs were also used to represent the data visually. For Data Cleaning ,Jupyter notebook is used with python along with pandas ,NumPy, Seaborn and matplotlib pakages.

Insights:

A. Cleaning the data :

Cleaning the data is one of the most important steps before moving forward with analysis. We dropped some columns that were not useful for our analysis and removed null values by droping, imputation with median and replacement .We also created a new column called profit which contained the difference between gross and budget.
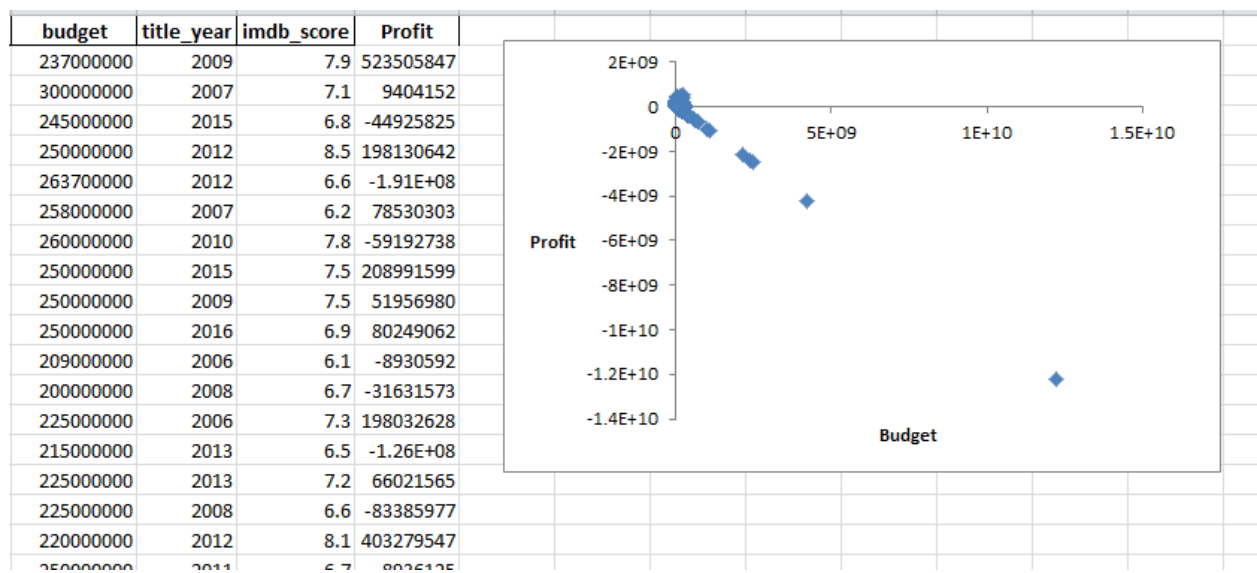
Remaining columns are:
'director_name', 'num_critic_for_reviews', 'gross', 'genres', 'actor_1_name', 'movie_title', 'num_voted_users', 'num_user_for_reviews', 'language', 'budget', 'title_year', 'imdb_score'

Detailed Data Cleaning Process is Provied in jupyter notebook python file.

B. Movies with highest profit:
We found the movies with the highest profit by sorting the dataset based on the profit column.

budget	title_year	imdb_score	Profit
237000000	2009	7.9	523505847
300000000	2007	7.1	9404152
245000000	2015	6.8	-44925825
250000000	2012	8.5	198130642
263700000	2012	6.6	-1.91E+08
258000000	2007	6.2	78530303
260000000	2010	7.8	-59192738
250000000	2015	7.5	208991599
250000000	2009	7.5	51956980
250000000	2016	6.9	80249062
209000000	2006	6.1	-8930592
200000000	2008	6.7	-31631573
225000000	2006	7.3	198032628
215000000	2013	6.5	-1.26E+08
225000000	2013	7.2	66021565
225000000	2008	6.6	-83385977
220000000	2012	8.1	403279547
250000000	2011	6.7	8926125

Avatar is the movie with Highest Profit .

C. Top 250:

Rank	IMDB_Top_250
1	The Shawshank Redemption
2	The Godfather
3	The Dark Knight
4	The Godfather: Part II
5	The Lord of the Rings: The Return of the King
6	Schindler's List
7	Pulp Fiction
8	The Good, the Bad and the Ugly
9	Inception
10	The Lord of the Rings: The Fellowship of the Ring
11	Fight Club
12	Forrest Gump
13	Star Wars: Episode V - The Empire Strikes Back
14	The Lord of the Rings: The Two Towers
15	The Matrix
16	Goodfellas
17	Star Wars: Episode IV - A New Hope
18	One Flew Over the Cuckoo's Nest
19	City of God
20	Seven Samurai
21	Interstellar
22	Saving Private Ryan
23	Se7en
24	The Silence of the Lambs
25	Spirited Away
26	American History X
27	The Usual Suspects
28	Modern Times
29	The Dark Knight Rises
30	Gladiator
31	Terminator 2: Judgment Day
32	Django Unchained
33	The Departed
34	The Lion King
35	The Green Mile
36	The Prestige
37	The Pianist

38	Apocalypse Now
39	Raiders of the Lost Ark
40	Psycho
41	Back to the Future
42	Alien
43	Memento
44	Whiplash
45	The Lives of Others
46	Children of Heaven
47	WALL·E
48	Braveheart
49	Amélie
50	Baahubali: The Beginning
51	Star Wars: Episode VI - Return of the Jedi
52	Once Upon a Time in America
53	Princess Mononoke
54	Aliens
55	American Beauty
56	Lawrence of Arabia
57	Das Boot
58	Requiem for a Dream
59	Oldboy
60	Reservoir Dogs
61	A Separation
62	Toy Story 3
63	Up
64	Inside Out
65	Batman Begins
66	Inglourious Basterds
67	Indiana Jones and the Last Crusade
68	L.A. Confidential
69	Toy Story
70	Scarface
71	Eternal Sunshine of the Spotless Mind
72	Amadeus
73	Raging Bull
74	Metropolis
75	Unforgiven
76	Downfall
77	Room
78	Snatch

79	2001: A Space Odyssey
80	Good Will Hunting
81	The Sting
82	The Hunt
83	Some Like It Hot
84	Monty Python and the Holy Grail
85	Captain America: Civil War
86	How to Train Your Dragon
87	The Wolf of Wall Street
88	Finding Nemo
89	A Beautiful Mind
90	Casino
91	V for Vendetta
92	The Thing
93	Die Hard
94	Blade Runner
95	Gran Torino
96	Warrior
97	Howl's Moving Castle
98	Into the Wild
99	Pan's Labyrinth
100	The Big Lebowski
101	Incendies
102	Gone with the Wind
103	The Secret in Their Eyes
104	Trainspotting
105	The Bridge on the River Kwai
106	Lock, Stock and Two Smoking Barrels
107	On the Waterfront
108	The Avengers
109	Guardians of the Galaxy
110	Mad Max: Fury Road
111	The Revenant
112	The Bourne Ultimatum
113	Pirates of the Caribbean: The Curse of the Black Pearl
114	Monsters, Inc.
115	The Martian
116	Shutter Island
117	Jurassic Park
118	Gone Girl
119	The Truman Show

120	Deadpool
121	Kill Bill: Vol. 1
122	Prisoners
123	The Sixth Sense
124	Sin City
125	Rush
126	The Grand Budapest Hotel
127	Million Dollar Baby
128	The Help
129	No Country for Old Men
130	There Will Be Blood
131	12 Years a Slave
132	Spotlight
133	Hotel Rwanda
134	The Imitation Game
135	The Princess Bride
136	Groundhog Day
137	The Sea Inside
138	Tae Guk Gi: The Brotherhood of War
139	Stand by Me
140	Akira
141	Elite Squad
142	The Terminator
143	Platoon
144	Butch Cassidy and the Sundance Kid
145	Donnie Darko
146	Annie Hall
147	The Wizard of Oz
148	Before Sunrise
149	The Best Years of Our Lives
150	Amores Perros
151	The Celebration
152	Rocky
153	X-Men: Days of Future Past
154	Ratatouille
155	Star Trek
156	Life of Pi
157	Casino Royale
158	Blood Diamond
159	The Incredibles
160	Cinderella Man

161	Big Fish
162	The Pursuit of Happyness
163	Kill Bill: Vol. 2
164	Catch Me If You Can
165	The Iron Giant
166	JFK
167	Serenity
168	Magnolia
169	District 9
170	Mystic River
171	Aladdin
172	Rain Man
173	Her
174	Dances with Wolves
175	Dead Poets Society
176	The Artist
177	The King's Speech
178	Brazil
179	In Bruges
180	Mulholland Drive
181	Slumdog Millionaire
182	Black Swan
183	The Perks of Being a Wallflower
184	True Romance
185	Dancer in the Dark
186	The Exorcist
187	Jaws
188	Casino Royale
189	Doctor Zhivago
190	The Straight Story
191	Fiddler on the Roof
192	Sicko
193	My Name Is Khan
194	The Sound of Music
195	Persepolis
196	Dallas Buyers Club
197	Shaun of the Dead
198	Sling Blade
199	Boyhood
200	Bowling for Columbine
201	Central Station

202	Young Frankenstein
203	Before Sunset
204	Waltz with Bashir
205	A Fistful of Dollars
206	Avatar
207	The Hobbit: The Desolation of Smaug
208	Iron Man
209	Edge of Tomorrow
210	Big Hero 6
211	The Hobbit: An Unexpected Journey
212	How to Train Your Dragon 2
213	Toy Story 2
214	Children of Men
215	The Insider
216	The Hateful Eight
217	The Bourne Identity
218	Almost Famous
219	Captain Phillips
220	Shrek
221	Hero
222	The Notebook
223	Glory
224	Walk the Line
225	Straight Outta Compton
226	The Blues Brothers
227	The Right Stuff
228	Taken
229	The Untouchables
230	The World's Fastest Indian
231	Edward Scissorhands
232	Glory
233	Ed Wood
234	My Fair Lady
235	Halloween
236	Hot Fuzz
237	Crouching Tiger, Hidden Dragon
238	The Remains of the Day
239	Boogie Nights
240	Letters from Iwo Jima
241	The Fighter
242	E.T. the Extra-Terrestrial

243	Crash
244	Amour
245	Nightcrawler
246	Little Miss Sunshine
247	Veer-Zaara
248	The Wrestler
249	The Chorus
250	Do the Right Thing

The Shawshank Redemption is the highest rated movie of all time.

D. Best Directors:

Director_name	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Richard Marquand	8.4
S.S. Rajamouli	8.4

S.S. Rajamouli is the only Indian director in the list of top 10 director.

E. Popular Genres:

Most Popular Genres	Average of imdb_score
Crime|Drama|Fantasy|Mystery	8.5
Adventure|Animation|Drama|Family|Musical	8.5
Adventure|Animation|Fantasy	8.4
Action|Adventure|Drama|Fantasy|War	8.4
Adventure|Drama|Thriller|War	8.4
Documentary|Drama|Sport	8.3
Documentary|War	8.3
Biography|Drama|History|Music	8.3
Adventure|Animation|Comedy|Drama|Family|Fantasy	8.3
Adventure|Drama|War	8.25

In the top 10 according to mean of imdb_score Drama is the most common ,Hence Drama is #1 in term of popular Genres.
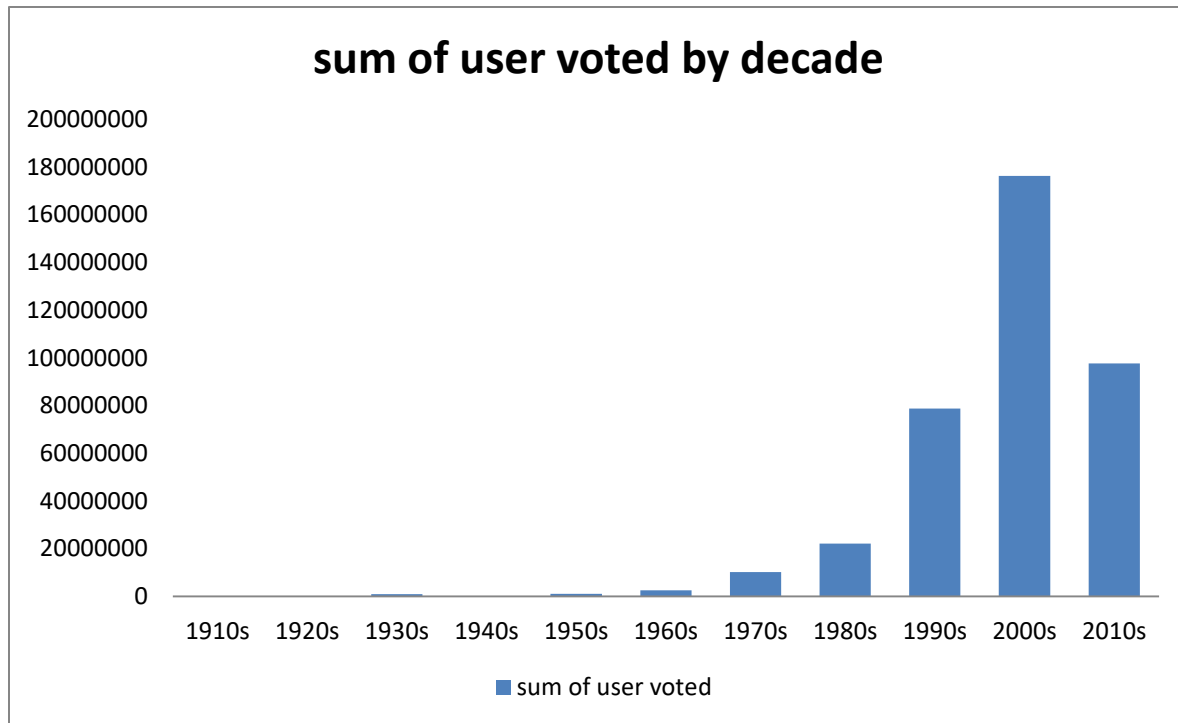
F. Charts:

Movies Extraction of all 3 actors

Meryl Streep	Leonardo DiCaprio	Brad Pitt
It's Complicated	Titanic	The Curious Case of Benjamin Button
The River Wild	The Great Gatsby	Troy
Julie & Julia	Inception	Ocean's Twelve
The Devil Wears Prada	The Revenant	Mr. & Mrs. Smith
Lions for Lambs	The Aviator	Spy Game
Out of Africa	Django Unchained	Ocean's Eleven
Hope Springs	Blood Diamond	Fury
One True Thing	The Wolf of Wall Street	Seven Years in Tibet
The Hours	Gangs of New York	Fight Club
The Iron Lady	The Departed	Sinbad: Legend of the Seven Seas
A Prairie Home Companion	Shutter Island	Interview with the Vampire: The Vampire Chronicles
	Body of Lies	The Tree of Life
	Catch Me If You Can	The Assassination of Jesse James by the Coward Robert Ford
	The Beach	Babel
	Revolutionary Road	By the Sea
	The Man in the Iron Mask	Killing Them Softly
	J. Edgar	True Romance
	The Quick and the Dead	
	Marvin's Room	
	Romeo + Juliet	
	The Great Gatsby	

From this we can say Leonardo Dicaprio play maximum lead role in movies out of all these actors.

Mean	Brad Pitt	Leonardo DiCaprio	Meryl Streep
Mean of num_user_for_reviews	742.3529412	914.4761905	297.1818182
Mean of num_critic_for_reviews	245	330.1904762	181.4545455
		Maximum Mean	

Based on mean of num_critic_for_reviews and mean of num_user_for_reviews we find that Leonardo DiCaprio is the critic-favorite and audience-favorite actor.

Below is the bar chart showing the number of users voted by decade .We can clearly see that maximum numbers of user voted in the decade 2000s.

sum of user voted by decade

Result:

In conclusion, we were able to derive valuable insights from the IMDB movie dataset. We framed a problem, cleaned the data, and used various data analysis techniques such as grouping, sorting, and plotting. We identified the factors contributing to a movie's success and found the top directors and popular genres. We also found the critic-favorite and audience-favorite actors and observed the change in the number of voted users over decades. Our analysis can help filmmakers and producers in making data-driven decisions.