IMDB Movie Analysis

: Final Project-1

By: Arjun P

Project Description:-

IMDB Movie Analysis project is about finding trends and insights about movies dataset from IMDB reviewing website. In this project, I have used the imdb_movie dataset provided by trainity and drawn some conclusions. I have provided insights to topics and answered the questions asked by the management team. I have used Google Spreadsheets and Microsoft Excel for data analytics and data visualization.

Approach:-

Firstly, I have used the basics of the data analytics process to clean the raw data and ask questions from cleaned data. Then, I have used data wrangling to make small data frames for relevant insights to answer all the possible questions. Finally, I combined all the results and visuals into this report.

Tech-Stack Used:-

I have used the web based application "**Google Sheets**" which is part of google online docs and "**Microsoft Excel for Mac version 16.70"** for performing various functions on spreadsheets. Both of these software provide ease of work and make data sharing and real time tracking very easy.

Project Insights:-

The database has only one table named IMDB_Movies:

Table Name	No. of Rows	No. of Columns
IMDB_Movies	5043	28

Table Details:

Column Name	Null Values	Description
color	19	Picture color of movie(Black and white or color)
director_name	103	Name of director of movie
num_critic_for_reviews	49	Number of critic reviews of movie
duration	15	Duration or length of the movie
director_facebook_likes	103	Likes on facebook handle of director of movie
actor_3_facebook_likes	23	Likes on facebook handle of actor_3 of movie
actor_2_name	13	Name of 2nd coactor
actor_1_facebook_likes	7	Likes on facebook handle of actor_1 of movie
gross	874	Gross revenue done by movie
genres	0	Different categories to which movie belongs
actor_1_name	7	Name of 1st actor
movie_title	0	Title of the movie
num_voted_users	0	Number of users voted for movie on imdb
cast_total_facebook_likes	0	Total likes on facebook handle of whole cast of movie
actor_3_name	23	Name of 3rd coactor
facenumber_in_poster	13	Number of faces in Movies Poster
plot _key words	152	Keywords referring to plot of movie
movie_imdb_link	0	Imdb link of movie
num_user_for_reviews	21	Number of users who reviewed the movie
language	12	Language of movie
countr y	5	Country to which movie belongs to
content_rating	301	Content rating of the movie
budget	487	Budget of Movie
title_year	107	The year in which movie is released
actor_2_facebook_likes	13	Likes on facebook handle of actor_2 of movie
imdb_score	0	IMDB score of movie on scale of (1-10)

aspect_ratio	327	Aspect ratio the movie made in
movie_facebook_likes	0	Likes on facebook handle of movie

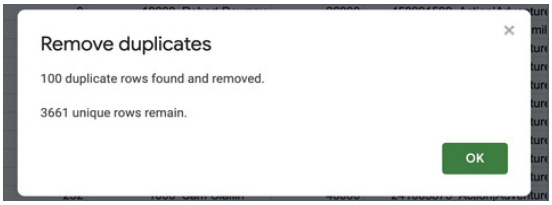
A.Cleaning the Data:-

Removing Null values:

- I used the =countif(A2:AB2,"")formulato checkthe numberof null valuesin a row then used the filter function to filter out 1287 rows with null values.
- Before deleting the rows, I checked for attributes which could have 0 as a value like director_facebook_likes,actor_3_facebook_likes and many more. After listing out these attributes I replaced the null values in these columns with 0. Then, the remaining 1282 rows with null values in other attributes which cannot be replaced like duration, director_name, and many more are removed.
- After removing null values 3761 rows of data is left which means around 26% data set included null values which is removed.

Removing Duplicates:

- In google sheet, there is a direct option to remove duplicates using the Data cleanup function present in the Data function in ribbon.
- Select Data>Data Cleanup>Remove Duplicates>Select Column which are constant in the movie data like color, director_name, duration, gross, genres, movie_title, plot_keywords, movie_imdb_link, language, country, budget, title_year, imdb_score.
- 100 duplicate rows are found and 3661 unique rows are left as shown in image below



Finding outliers:

- In the country column at row 2391 the value is the official site which is an error. As it was a single error in the column full of country names the value was corrected using data from movie_imdb_link to USA.
- In the budget column at row 2642 for movie name “The Host” there is an outlier like 12 billion usd for movie budget which is not possible so it would probably be in south korean currency and is removed. Also there is a possibility that some movies from other countries could have the wrong

budget or gross information. Some of them are removed including Lady Vengeance, Fateless, Princess Mononoke, Steamboy, Akira, Godzilla 2000, Tango, Kabhi Alvida Naa Kehna, and Red Cliff leaving behind 3652 rows. But all of those cannot be checked efficiently so either we can ignore or remove this data. Around 80% of the data is from the USA so it is best to ignore other outliers in this column.

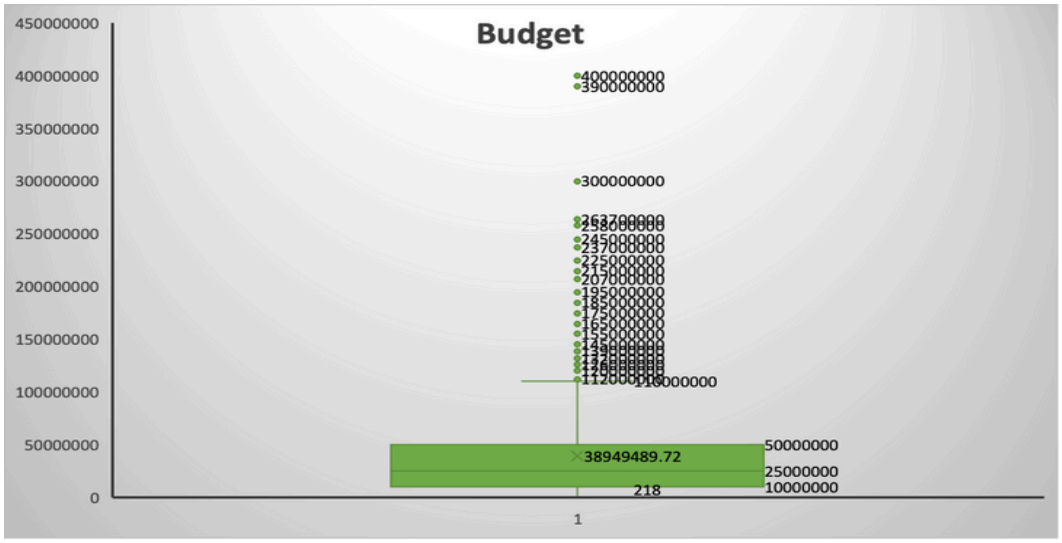


Fig 1: Box plot for outliers in Budget column

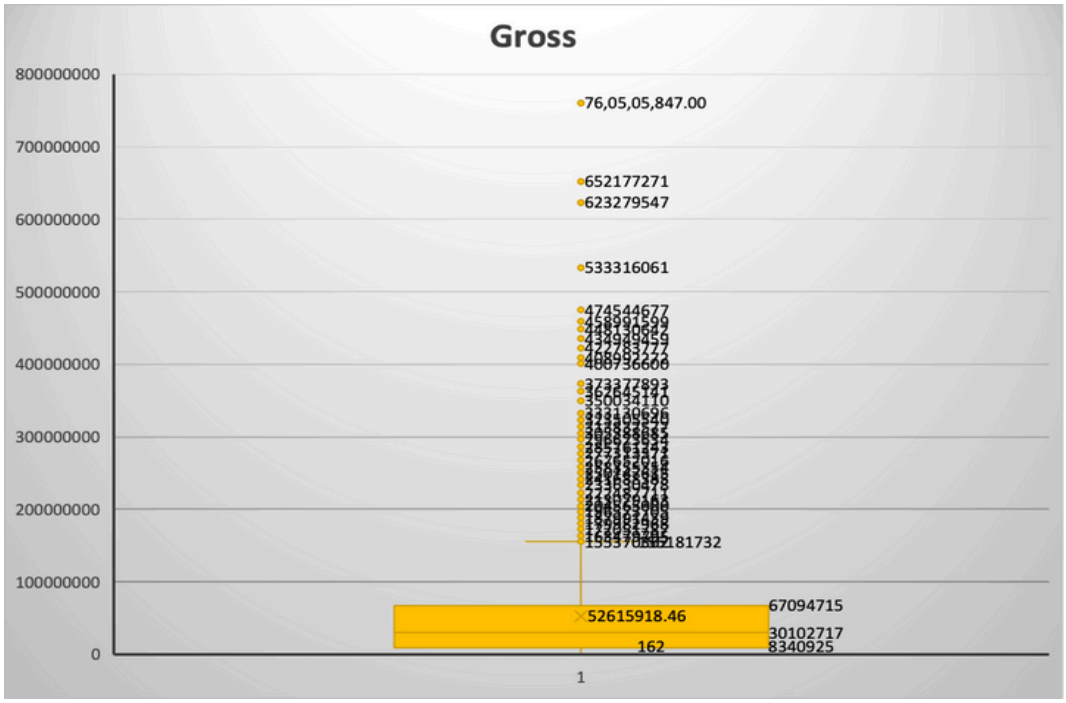


Fig 2: Box plot for outliers in Gross column

B.Movies with highest profit:

Find the movies with the highest profit?

Approach- I created a new column called profit which contains the difference of the two columns: gross and budget. Then, sorted the column using the profit column as reference and Plotted profit (y-axis) vs budget (x- axis) in scatter plot. Also, observed the outliers using the Histogram chart type. To find movies with highest profit i have sorted the data set using Profit in Descending order and extracted the table which is represented in the column chart.

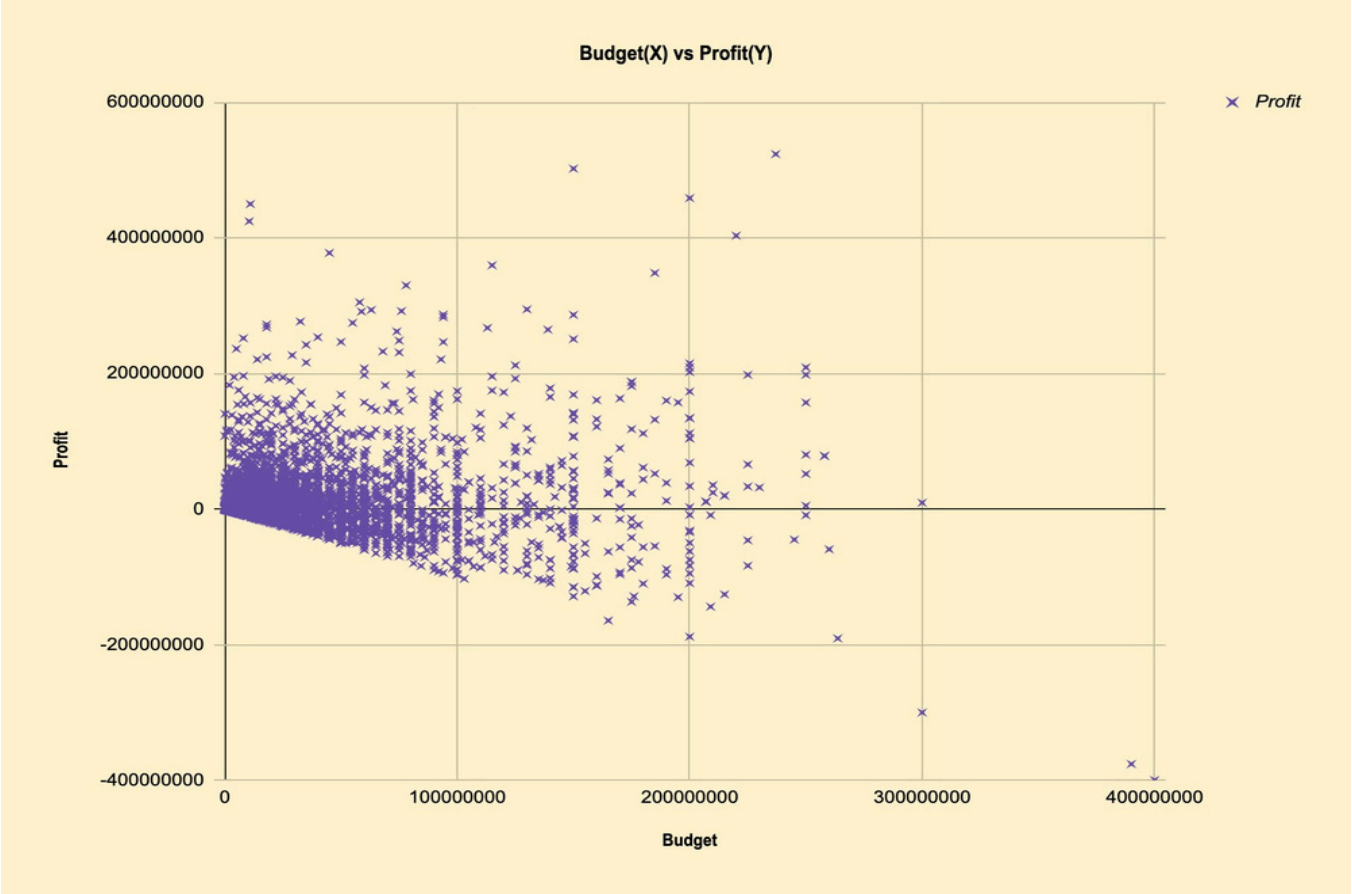


Fig 3: Scatter Plot for Budget vs Profit where x denotes the data points on axis and data point further from cluster are outliers

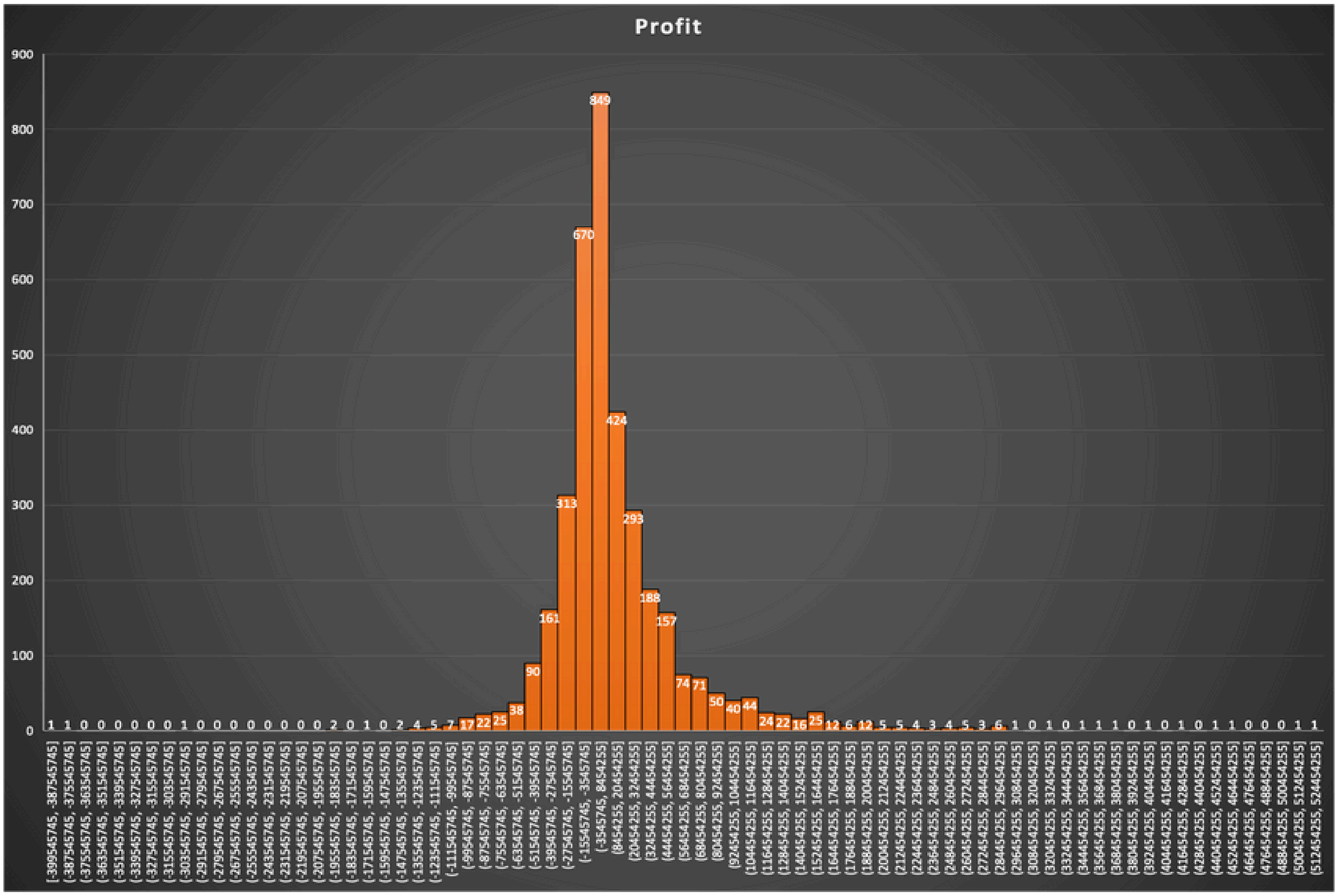


Fig 4: Histogram Representing distribution of profit and outlier in the column

Table 1: Top 10 movies which made Highest profit

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

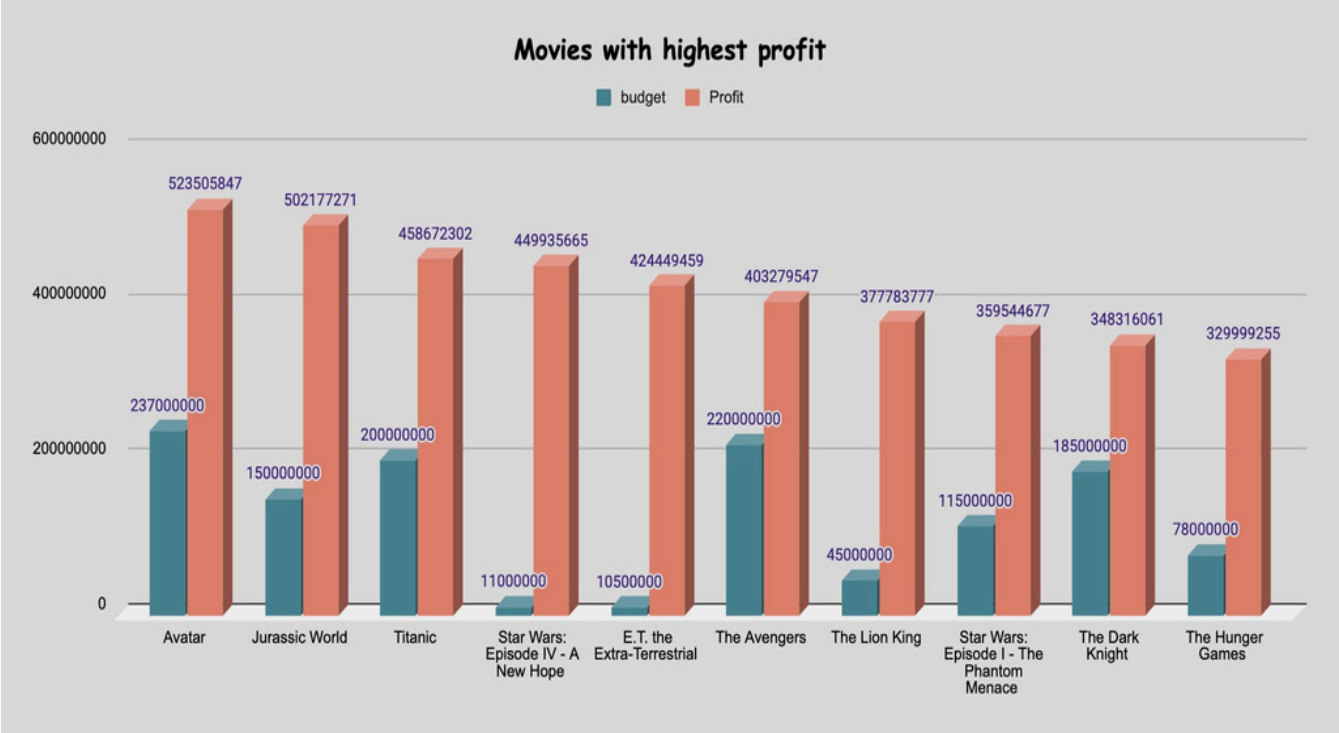


Fig 5: Top 10 movies which made Highest profit

Conclusion :- The top 10 movies which made the highest profit can be seen in table and chart representation.

C.Top 250

Approach - I created a new table IMDb_Top_250 and then stored the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also used filter by condition according to the num_voted_users column where value is greater than 25,000. Also added a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films. Then, all the movies in the IMDb_Top_250 column which are not in the English language are extracted into new column named Top_Foreign_Lang_Film.

Result Grid :-IMDb_Top_250

Rank	movie_title	num_voted_users	imdb_score	Rank	movie_title	num_voted_users	imdb_score
1	The Shawshank Redemption	1689764	9.3	126	The Grand Budapest Hotel	475518	8.1
2	The Godfather	1155770	9.2	127	The Martian	472488	8.1
3	The Dark Knight	1676169	9 9	128	The Imitation Game	467613	8.1
4	The Godfather: Part II	790926	8.9	129	12 Years a Slave	439176	8.1
5	Pulp Fiction	1324680		130	Groundhog Day	437418	8.1
6	The Lord of the Rings: The Return of the King	1215718	8.9	131	The Revenant	406020	8.1
7	Schindler's List	865020	8.9	132	Prisoners	383591	8.1
8	The Good, the Bad and the Ugly	503509	8.9	133	Rocky	375240	8.1
9	Inception	1468200	8.8	134	There Will Be Blood	372990	8.1

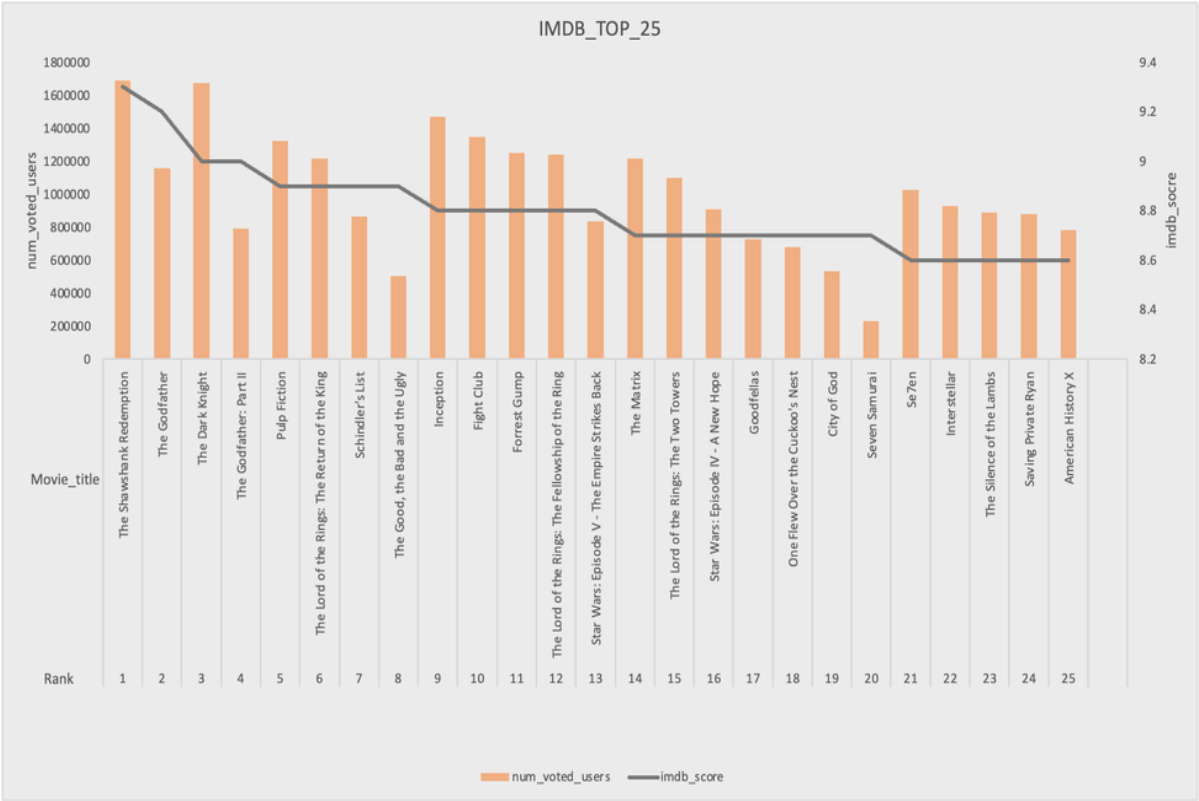
Rank	movie_title	num_voted_users	imdb_score	Rank	movie_title	num_voted_users	imdb_score
10	Fight Club	1347461	8.8	135	The Help	318955	8.1
11	Forrest Gump	1251222	8.8	136	Rush	312629	8.1
12	The Lord of the Rings: The Fellowship of the Ring	1238746	8.8	137	The Princess Bride	294163	8.1
13	Star Wars: Episode V - The Empire Strikes Back	837759	8.8	138	The Wizard of Oz	291875	8.1
14	The Matrix	1217752	8.7	139	Platoon	291603	8.1
15	The Lord of the Rings: The Two Towers	1100446	8.7	140	Stand by Me	271794	8.1
16	Star Wars: Episode IV - A New Hope	911097	8.7	141	Hotel Rwanda	264533	8.1
17	Goodfellas	728685	8.7	142	Spotlight	195333	8.1
18	One Flew Over the Cuckoo's Nest	680041	8.7	143	Annie Hall	192940	8.1
19	City of God	533200	8.7	144	Before Sunrise	183288	8.1
20	Seven Samurai	229012	8.7	145	Amores Perros	173551	8.1
21	Se7en	1023511	8.6	146	Butch Cassidy and the Sundance Kid	152089	8.1
22	Interstellar	928227	8.6	147	Elite Squad	81644	8.1
23	The Silence of the Lambs	887467	8.6	148	The Celebration	65951	8.1
24	Saving Private Ryan	881236	8.6	149	The Sea Inside	64556	8.1
25	American History X	782437	8.6	150	The Best Years of Our Lives	40359	8.1
26	The Usual Suspects	740918	8.6	151	Tae Guk Gi: The Brotherhood of War	31943	8.1
27	Spirited Away	417971	8.6	152	Woodstock	12631	8.1
28	Modern Times	143086	8.6	153	In the Shadow of the Moon	5475	8.1
29	The Dark Knight Rises	1144337	8.5	154	Slumdog Millionaire	641997	8
30	Gladiator	982637	8.5	155	Black Swan	551363	8
31	Django Unchained	955174	8.5	156	District 9	531737	8
32	The Departed	873649	8.5	157	Catch Me If You Can	525801	8
33	Memento	845580	8.5	158	X-Men: Days of Future Past	514125	8
34	The Prestige	844052	8.5	159	Kill Bill: Vol. 2	512749	8
35	The Green Mile	782610	8.5	160	Star Trek	504419	8
36	Terminator 2: Judgment Day	744891	8.5	161	The King's Speech	503631	8
37	Back to the Future	732212	8.5	162	The Incredibles	479166	8
38	Raiders of the Lost Ark	661017	8.5	163	Ratatouille	473887	8
39	The Lion King	644348	8.5	164	Casino Royale	470483	8
40	Alien	563827	8.5	165	Life of Pi	440084	8
41	The Pianist	497946	8.5	166	Jaws	412454	8
42	Apocalypse Now	450676	8.5	167	Blood Diamond	400292	8
43	Psycho	422432	8.5	168	Shaun of the Dead	395921	8
44	Whiplash	399138	8.5	169	Rain Man	383784	8

Rank	movie_title	num_voted_users	imdb_score	Rank	movie_title	num_voted_users	imdb_score
45 46 47 48 49	The Lives of Others	259379	8.5	170	Her	355126	8 8 8
	Children of Heaven	27882	8.5	171	The Perks of Being a Wallflower	351274	8 8
	American Beauty	822500	8.4	172	Big Fish	350698	
	Braveheart	736638	8.4	173	Mystic River	338415	
	WALL·E	718837	8.4	174	The Pursuit of Happyness	338383	
50	Star Wars: Episode VI - Return of the Jedi	681857	8.4	175	Dallas Buyers Club	326494	8
51	Reservoir Dogs	664719	8.4	176	In Bruges	307639	8
52	Requiem for a Dream	573541	8.4	177	The Exorcist	284252	8
53	Amélie	534262	8.4	178	Dead Poets Society	277451	8
54	Aliens	488537	8.4	179	Boyhood	266020	8
55	Oldboy	356181	8.4	180	Aladdin	260939	8
56	Once Upon a Time in America	221000	8.4	181	Serenity	242599	8
57	Lawrence of Arabia	192775	8.4	182	Magnolia	241030	8
58	Das Boot	168203	8.4	183	Mulholland Drive	235992	8
59	A Separation	151812	8.4	184	The Artist	190030	8
60	Batman Begins	980946	8.3	185	Dances with Wolves	186485	8
61	Inglourious Basterds	885175	8.3	186	Before Sunset	168398	8
62	Eternal Sunshine of the Spotless Mind	666937	8.3	187	True Romance	163492	8
63	Up	665575	8.3	188	Brazil	152306	8
64	Toy Story	623757	8.3	189	Cinderella Man	148238	8
65	Good Will Hunting	604904	8.3	190	The Sound of Music	148172	8
66	Snatch	600996	8.3	191	A Fistful of Dollars	147566	8
67	Toy Story 3	544884	8.3	192	The Iron Giant	128455	8
68	Scarface	537442	8.3	193	Bowling for Columbine	123090	8
69	Indiana Jones and the Last Crusade	515306	8.3	194	JFK	113472	8
70	2001: A Space Odyssey	427357	8.3	195	Young Frankenstein	112671	8
71	L.A. Confidential	414219	8.3	196	Dancer in the Dark	79330	8
72	Monty Python and the Holy Grail	382240	8.3	197	Sling Blade	72443	8
73	Inside Out	345198	8.3	198	Persepolis	70194	8
74	Unforgiven	277505	8.3	199	My Name Is Khan	69759	8
75	Amadeus	270790	8.3	200	Sicko	66610	8
76	Downfall	248354	8.3	201	The Straight Story	63733	8
77	Raging Bull	235133	8.3	202	Doctor Zhivago	55816	8
78	The Sting	175607	8.3	203	Waltz with Bashir	46107	8
79	Some Like It Hot	175196	8.3	204	Fiddler on the Roof	29839	8
80	The Hunt	170155	8.3	205	Central Station	28951	8

Rank	movie_title	num_voted_users	imdb_score	Rank	movie_title	num_voted_users	imdb_score
81 82 83	Room	161288	8.3	206	Blood In, Blood Out	23181	8 7.9
	Metropolis	111841	8.3	207	Avatar	886204	7.9
	Hoop Dreams	18980	8.3	208	Iron Man	696338	
84	V for Vendetta	791783	8.2	209	The Hobbit: An Unexpected Journey	637246	7.9
85	The Wolf of Wall Street	780588	8.2	210	Taken	483756	7.9
86	Finding Nemo	692482	8.2	211	The Hobbit: The Desolation of Smaug	483540	7.9
87	A Beautiful Mind	610568	8.2	212	Shrek	467113	7.9
88	Die Hard	592582	8.2	213	Edge of Tomorrow	431620	7.9
89	Gran Torino	561773	8.2	214	The Bourne Identity	407601	7.9
90	The Big Lebowski	537419	8.2	215	The Notebook	396396	7.9
91	How to Train Your Dragon	485430	8.2	216	Toy Story 2	385871	7.9
92	Trainspotting	469561	8.2	217	Children of Men	361767	7.9
93	Pan's Labyrinth	467234	8.2	218	Crash	361169	7.9
94	Blade Runner	461609	8.2	219	Edward Scissorhands	357581	7.9
95	Into the Wild	426359	8.2	220	Little Miss Sunshine	355810	7.9
96	Lock, Stock and Two Smoking Barrels	414976	8.2	221	Hot Fuzz	352695	7.9
97	Casino	333542	8.2	222	Captain Phillips	323353	7.9
98	Warrior	332276	8.2	223	Nightcrawler	293304	7.9
99	Captain America: Civil War	272670	8.2	224	E.T. the Extra-Terrestrial	281842	7.9
100	The Thing	258078	8.2	225	Big Hero 6	279093	7.9
101	Gone with the Wind	215340	8.2	226	The Fighter	275869	7.9
102	Howl's Moving Castle	214091	8.2	227	The Hateful Eight	272839	7.9
103	The Bridge on the River Kwai	149444	8.2	228	Moon	260607	7.9
104	The Secret in Their Eyes	131831	8.2	229	The Wrestler	251349	7.9
105	On the Waterfront	100890	8.2	230	How to Train Your Dragon 2	221128	7.9
106	Incendies	80429	8.2	231	The Untouchables	219008	7.9
107	The Act of Killing	23836	8.2	232	Crouching Tiger, Hidden Dragon2	17740	7.9
108	The Avengers	995415	8.1	233	Almost Famous	207287	7.9
109	Pirates of the Caribbean: The Curse of the Black Pearl	809474	8.1	234	Boogie Nights	189032	7.9
110	Shutter Island	786092	8.1	235	Walk the Line	188637	7.9
111	Kill Bill: Vol. 1	735784	8.1	236	Halloween	157857	7.9
112	The Sixth Sense	704766	8.1	237	Hero	149414	7.9
113	Guardians of the Galaxy	682155	8.1	238	The Blues Brothers	142448	7.9
114	The Truman Show	667983	8.1	239	Ed Wood	142416	7.9
115	Sin City	656640	8.1	240	The Insider	133526	7.9
116	Jurassic Park	613473	8.1	241	Letters from Iwo Jima	132149	7.9

Rank	movie_title	num_voted_users	imdb_score	Rank	movie_title	num_voted_users	imdb_score
117	No Country for Old Men	612060	8.1	242	Straight Outta Compton	119928	7.9
118	The Terminator	600266	8.1	243	Glory	101888	7.9
119	Monsters, Inc.	585659	8.1	244	Before Midnight	95362	7.9
120	Donnie Darko	580999	8.1	245	Once	90827	7.9
121	Gone Girl	569841	8.1	246	Amour	70382	7.9
122	Mad Max: Fury Road	552503	8.1	247	My Fair Lady	66959	7.9
123	The Bourne Ultimatum	491077	8.1	248	Do the Right Thing	59524	7.9
124	Million Dollar Baby	482064	8.1	249	The Remains of the Day	45703	7.9
125	Deadpool	479047	8.1	250	The Right Stuff	45271	7.9

Chart representation :-For top 25 movies accordingto IMDB rank



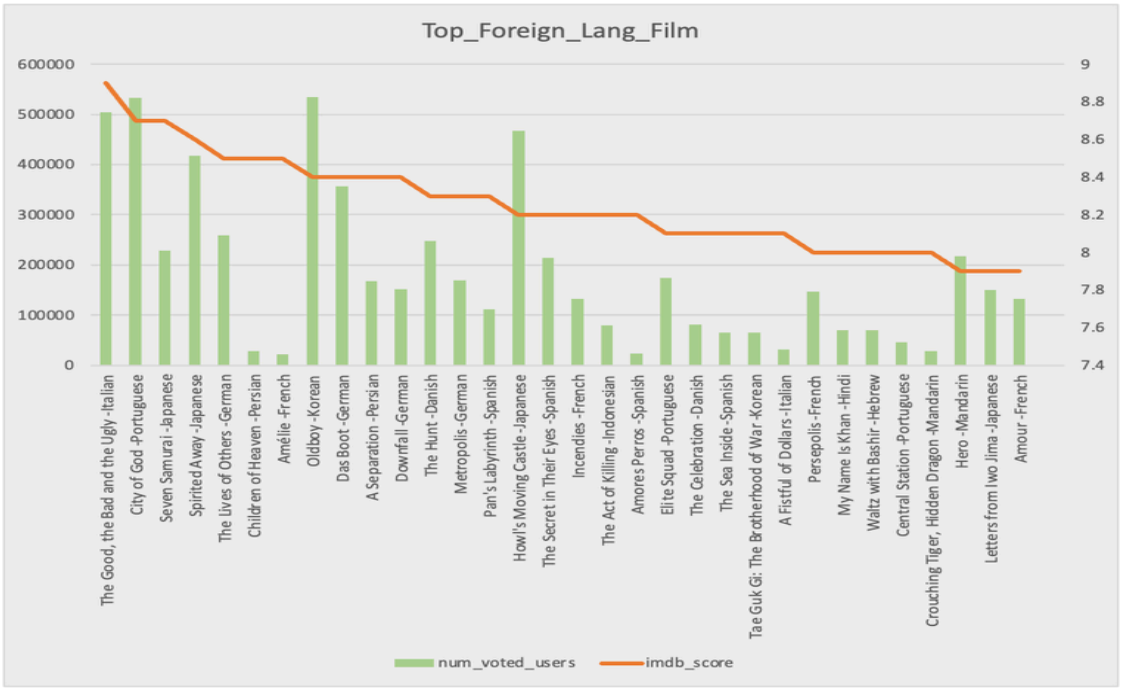
Conclusion :- The top 250 movies are shown in the result grid and chart representation for top 25 is shown. “The Shawshank Redemption” movie tops the list with an imdb score of 9.3.

Result Grid :-Top_Foreign_Lang_Film

Rank	movie_title	num_voted_users	language	country	imdb_score
8	The Good, the Bad and the Ugly	503509 533200	Italian	Italy	8.9 8.7 8.7
19	City of God	229012 417971	Portuguese	Brazil	8.6 8.5 8.5
20	Seven Samurai	259379 27882	Japanese	Japan	8.4
27	Spirited Away	534262	Japanese	Japan	
45	The Lives of Others		German	Germany	
46	Children of Heaven		Persian	Iran	
53	Amélie		French	France	

55	Oldboy	356181	Korean	South Korea	8.4
58	Das Boot	168203	German	West	8.4
59	A Separation	151812	Persian	Germany Iran	8.4
76	Downfall	248354	German	Germany	8.3
80	The Hunt	170155	Danish	Denmark	8.3
82	Metropolis	111841	German	Germany	8.3
93	Pan's Labyrinth	467234	Spanish	Spain Japan	8.2
102	Howl's Moving Castle	214091	Japanese	Argentina	8.2
104	The Secret in Their Eyes	131831	Spanish	Canada UK	8.2
106	Incendies	80429	French	Mexico Brazil	8.2
107	The Act of Killing	23836	Indonesian	Denmark	8.2
145	Amores Perros	173551	Spanish	Spain South	8.1
147	Elite Squad	81644	Portuguese	Korea Italy	8.1
148	The Celebration	65951	Danish	France India	8.1
149	The Sea Inside	64556	Spanish	Israel Brazil	8.1
151	Tae Guk Gi: The Brotherhood of War	31943	Korean	Taiwan China	8.1
191	A Fistful of Dollars	147566	Italian	USA France	8 8
198	Persepolis	70194	French		8 8
199	My Name Is Khan	69759	Hindi		8
203	Waltz with Bashir	46107	Hebrew		7.9
205	Central Station	28951	Portuguese		7.9
232	Crouching Tiger, Hidden Dragon	217740	Mandarin		7.9
237	Hero	149414	Mandarin		7.9
241	Letters from Iwo Jima	132149	Japanese		
246	Amour	70382	French		

Chart representation :-



Conclusion :- There are 32 foreign language films in IMDB_top_250 with IMDB scores ranging from 7.9 to 8.9.

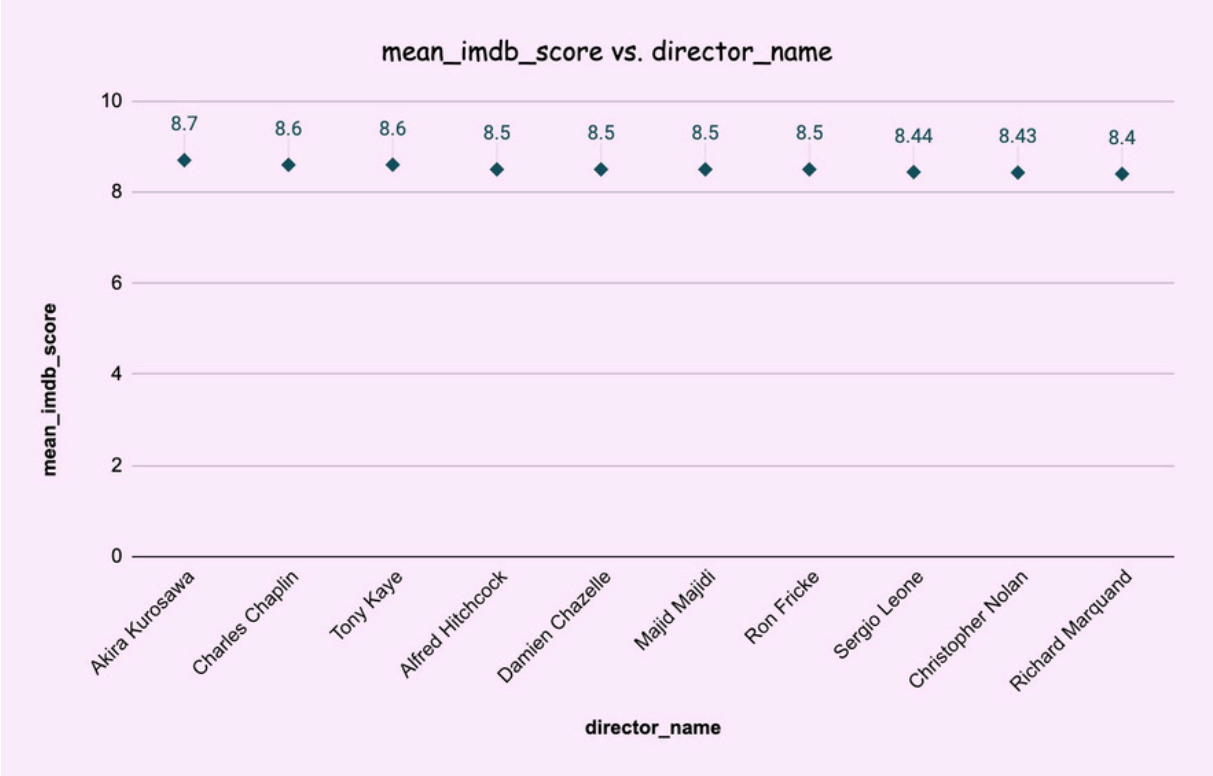
D.Best Directors:

Approach- I have used pivot table to extract director_namein rows from the dataset then used average of imdb score as values. Then used sort by to order data by mean imdb_score. Then, I used a custom sort function to sort the list first by mean imdb_score from Largest to smallest and then by director_name from A to z.

Result Grid :-

director_name	mean imdb_score
Akira Kurosawa	8.7
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	8.44
Christopher Nolan	8.43
Richard Marquand	8.4

Chart representation :-



Conclusion :- The top 10 director list is shown in the result grid as well as in chart representation. Director Akira Kurosawa tops the list with 8.7 imdb score.

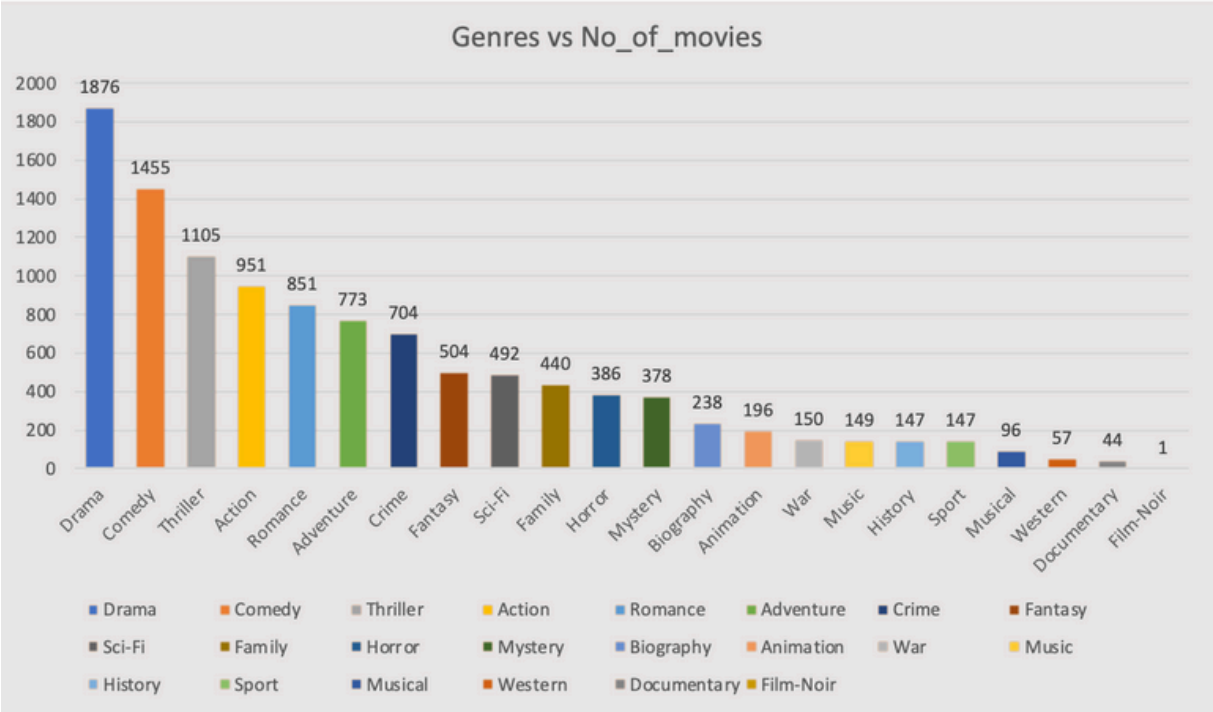
E.Popular Genres:

Approach- As the movies belong to multiple genres,I have used split text to column function to split genres column. Then I used the “unique” function to check the unique genres to which different movies belong. Then I used the “countif” function to check the number of movies with the said genres.

Result Grid :-

unique_genres	No_of_movies	% of Movie	unique_genres	% of Movie	No_of_movies
Drama	1876	50.39%	Mystery	10.15%	378
Comedy	1455	39.08%	Biography	6.39%	238
Thriller	1105	29.68%	Animation	5.26%	196
Action	951	25.54%	War	4.03%	150
Romance	851	22.86%	Music	4.00%	149
Adventure	773	20.76%	History	3.95%	147
Crime	704	18.91%	Sport	3.95%	147
Fantasy	504	13.54%	Musical	2.58%	96
Sci-Fi	492	13.22%	Western	1.53%	57
Family	440	11.82%	Documentary	1.18%	44
Horror	386	10.37%	Film-Noir	0.03%	1

Chart representation :-



Conclusion :- The most popular genre to which the majority of movies that is around 50% belong is Drama. Comedy and thriller genres made to the top 3 list.

F. Charts:

Find the critic-favorite and audience-favoriteactors

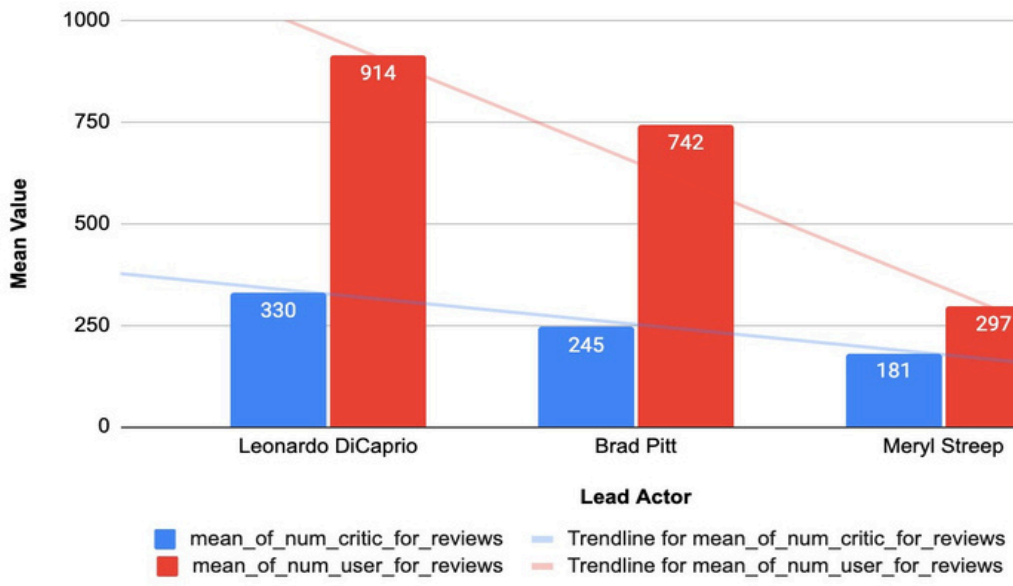
Approach- I created three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors. I used the actor_1_name column for extraction. Then, I appended the rows of all these columns and stored them in a new column named Combined. Group the combined column using the actor_1_name column. To observe the change in number of voted users over decades using a bar chart, I used a pivot table to group the title year into decades and corresponding to the sum of num_voted_users. Then filter it for meryl_streep, leo_caprio and brad_pitt.

Result Grid :-

1. Mean of the num_critic_for_reviews and num_users_for_reviewfor 3 actors

actor_1_name	mean_of_num_critic_for_reviews	mean_of_num_user_for_reviews
Leonardo DiCaprio	330 245 181	914 742 297
Brad Pitt		
Meryl Streep		

Chart representation :-

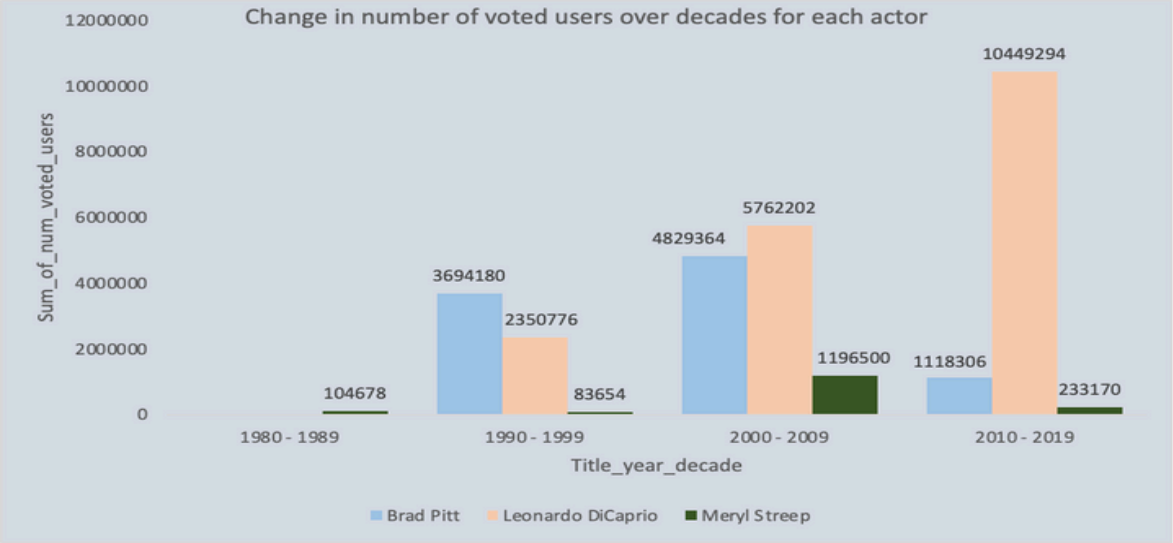


2. Change in number of voted users over decades for each actor

Sum_of_num_voted_users	*actor_1_name*		
Grouped title_year	Brad Pitt	Leonardo DiCaprio	Meryl Streep
1980 - 1989 1990 - 1999			104678
2000 - 2009	3694180	2350776	83654
	4829364	5762202	1196500

2010 - 2019	1118306	10449294	233170

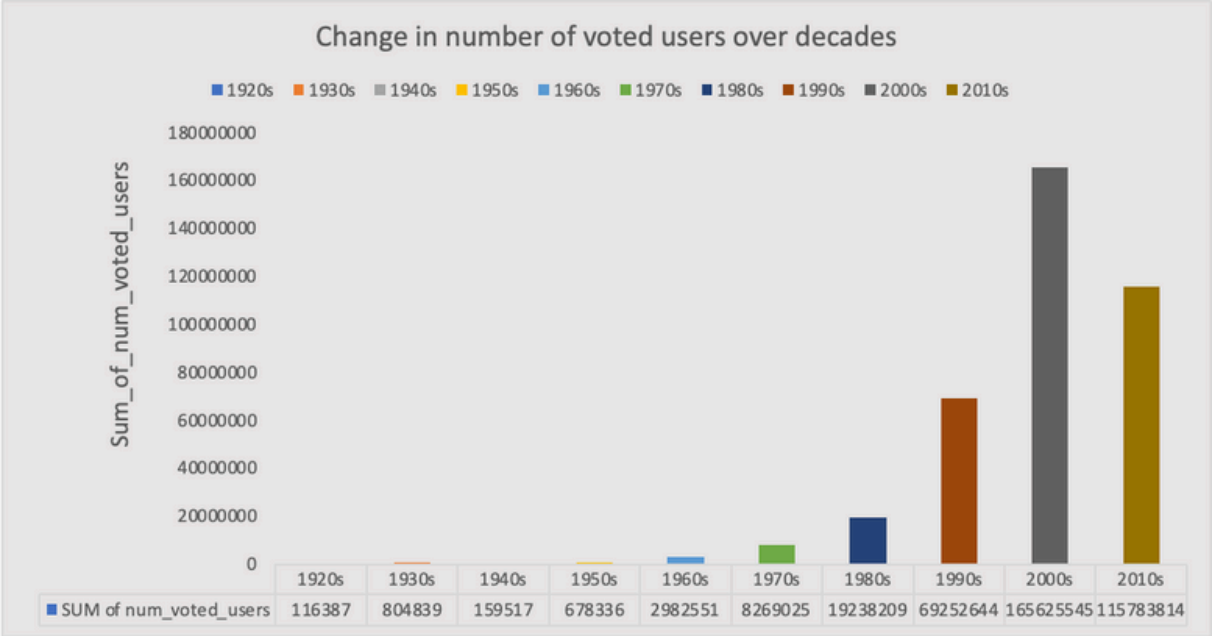
Chart representation :-



3. Change in number of voted users over decades for all movies in dataset

Grouped title_year	SUM of num_voted_users
1920s	116387
1930s	804839
1940s	159517
1950s	678336
1960s	2982551
1970s	8269025
1980s	19238209
1990s	69252644
2000s	165625545
2010s	115783814

.Chart representation :-



Conclusion:Leonardo DiCaprio is critic favorite anduser favorite actor with highest mean of the num critic for reviews and num users for review. Out of 3 actors, Leonardo DiCaprio has the highest number of voted users and has seen growth for 2 decades from 2000 to 2019. On the other hand, Brad Pitt had the highest number of voted users for 1 decade from 1990 - 1999 but the number declined in the decade of 2010s. The number of voted users has seen continuous growth from 1940s to 2010s. The possible reason for the decline in number for the last decade is due to the data being limited to half the decade.

Result:-

I have answered all the questions asked by the company in this project and explained the result grid and conclusion under the project insights part. While doing the project I applied my learning of statistics and understanding of different functions, pivot tables, conditionals used in spreadsheets.