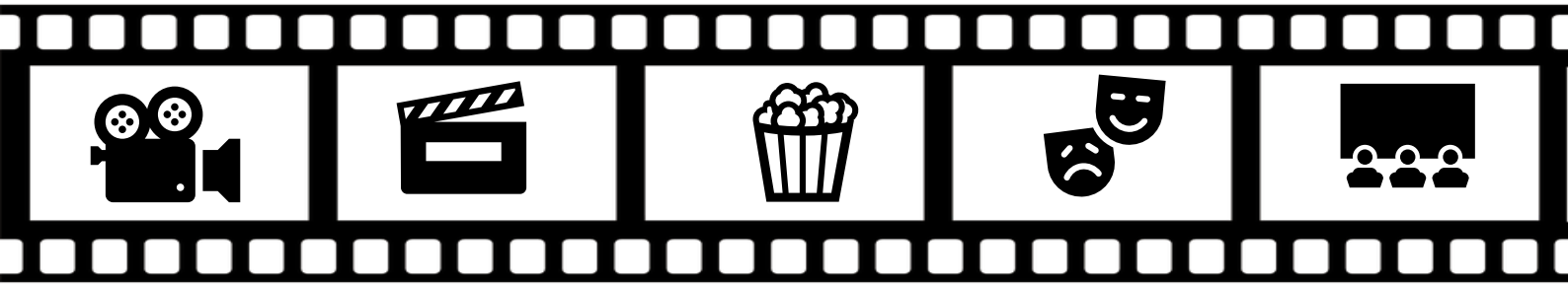


IMDB MOVIE ANALYSIS

Authored By,

Sushant Karmakar

INDEX

1. PROJECT DESCRIPTION

IMDb (an acronym for Internet Movie Database) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

As of March 2022, the database contained some 10.1 million titles (including television episodes) and 11.5-million-person records.

In this project, I have been provided with the dataset having various columns of different IMDB movies. I have been told to Frame the problem. Once I've defined a problem, it is suggested to clean the data as necessary and use Data Analysis skills to explore the data set and derive insights. Also make sure to use the process known as Root Cause Analysis, developed by Sakichi Toyoda, founder of Toyota Industries.

2. PROBLEMS & PLANNING

- **Cleaning the data:** This is one of the most important step to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data

- **Movies with highest profit:** Create a new column called profit which contains the difference of the two columns: gross and budget. Sort the column using the profit column as 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.

- **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 Language Film. You can use your own imagination also.

Your task: Find IMDB Top 250.

- **Best Directors:** Group 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 top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

Your task: Find the best directors.

- **Popular Genres:** Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres.

- **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-favourite and audience-favourite actors

3. PREPARATION & APPROACH

For this project, I have been provided a dataset in the form of csv file. It contains 5043 rows including every single detail about a movie. Before diving into analysis, it is required to clean and sort the data as per requirement. Then step by step, I'll be trying to solve the tasks that has been provided. After that, using the analysis data, I'll be using the Root-cause analysis process and figure out the important questions that needed to be answered. Also, the dataset has passed the ROCCC compliance process, hence making it a reliable one.

4. PROCESS & ANALYSIS

As the dataset has already been provided by the company, therefore carrying out the analysis will be way easier.

Task 1: Cleaning the data.

- At first, to evade data corruption, I made a copy of the raw data so that the original data won't get effected.
- Next, I have dropped off some unnecessary columns that won't be required for analysis. I have listed the columns below,

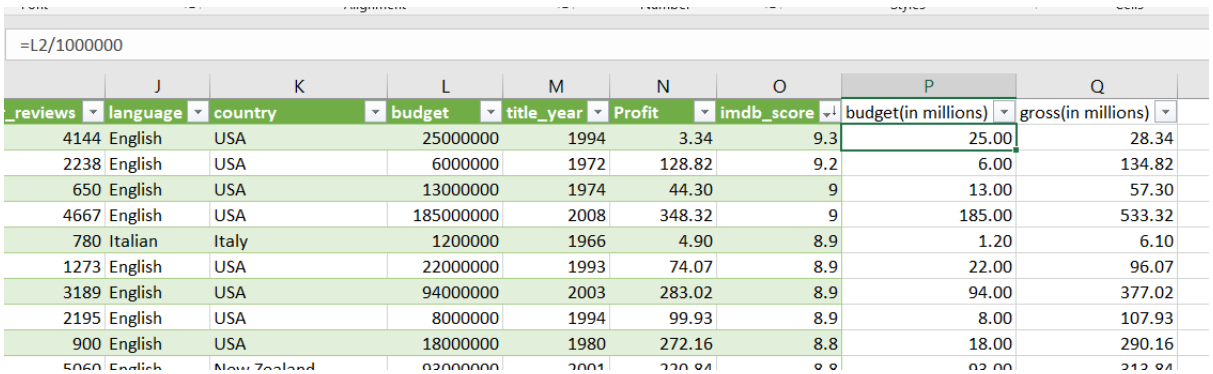
'colour', 'director_facebook_likes', 'actor_3_facebook_likes', 'actor_2_name', 'actor_1_facebook_likes', 'cast_total_facebook_likes', 'actor_3_name', 'facenumber_in_poster', 'plot_keywords', 'movie_imdb_link', 'content_rating', 'actor_2_facebook_likes', 'aspect_ratio' and 'movie_facebook_likes'.

- Next, I have removed the unnecessary rows as well only if it has any NULL values. Since gross and budget column had the most number of null values, therefore I filtered the blank values and deleted the whole row to avoid any miscalculations for further analysis.
- The language column had 3 empty cells, hence replacing them with English as they all belong to USA.
- After that, I have checked for any duplicate values in the dataset. This has been done using 'Remove Duplicate Value' from the Data tab.
- Therefore, after cleaning the table had 3784 rows along with 14 columns.
- The cleaned data can be seen on the ***IMDB Data(cleaned).xlsx*** under the name ***Cleaned Data.***

Task 2: Find the movie with highest profit.

To find the highest profit movie I did the following steps,

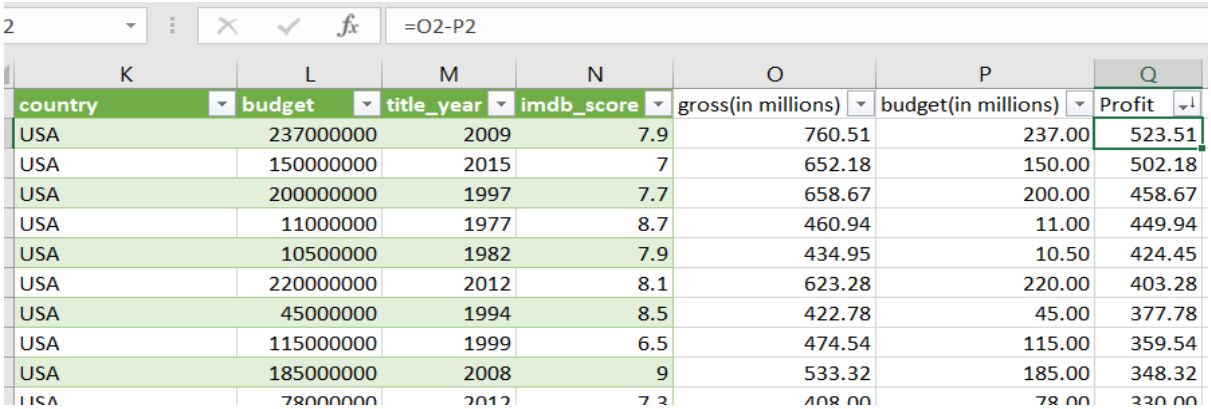
- I created two new columns and converted the budget and gross unit into millions by dividing them with 1000000.



=L2/1000000

J	K	L	M	N	O	P	Q
_reviews	language	country	budget	title_year	Profit	imdb_score	budget(in millions)	gross(in millions)
4144	English	USA	25000000	1994	3.34	9.3	25.00	28.34
2238	English	USA	6000000	1972	128.82	9.2	6.00	134.82
650	English	USA	13000000	1974	44.30	9	13.00	57.30
4667	English	USA	185000000	2008	348.32	9	185.00	533.32
780	Italian	Italy	1200000	1966	4.90	8.9	1.20	6.10
1273	English	USA	22000000	1993	74.07	8.9	22.00	96.07
3189	English	USA	94000000	2003	283.02	8.9	94.00	377.02
2195	English	USA	8000000	1994	99.93	8.9	8.00	107.93
900	English	USA	18000000	1980	272.16	8.8	18.00	290.16
5060	English	New Zealand	93000000	2001	220.84	8.8	93.00	313.84

- Next, I created a new column that depicts the difference between gross and budget and named it as profit.



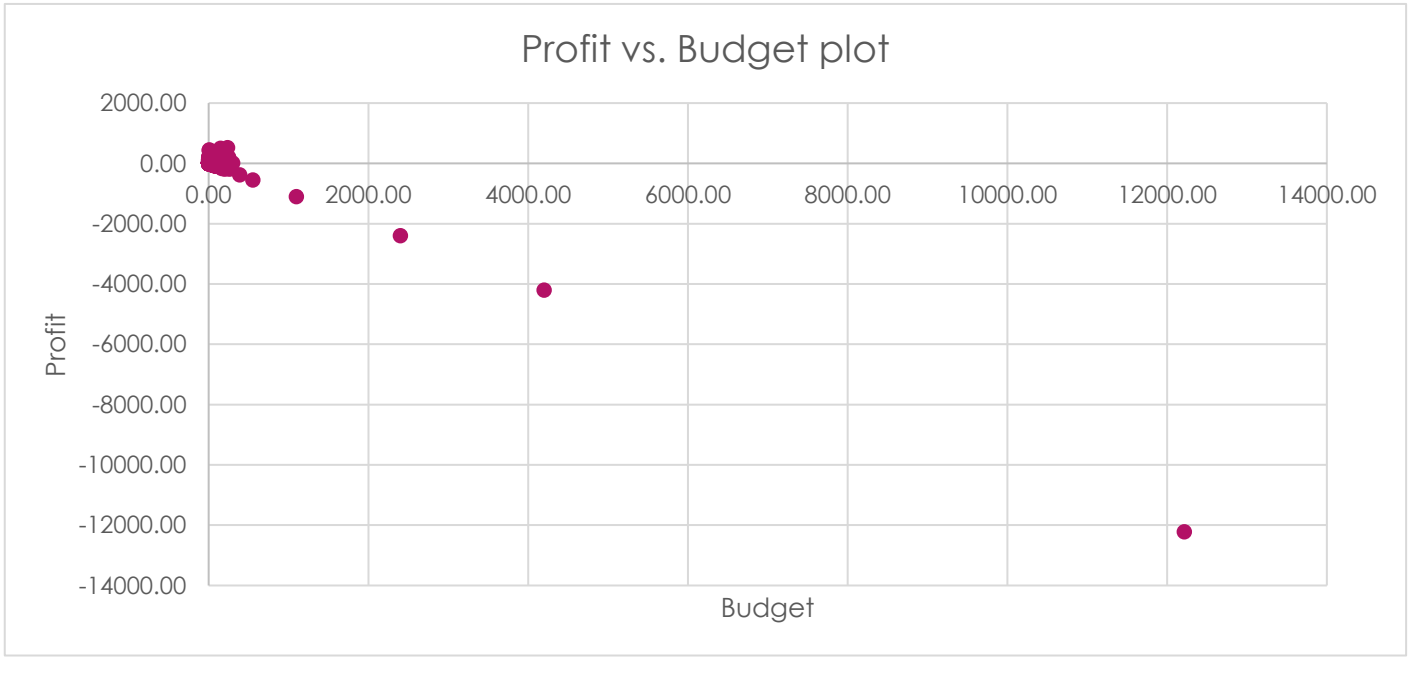
2 | =O2-P2

	K	L	M	N	O	P	Q
	country	budget	title_year	imdb_score	gross(in millions)	budget(in millions)	Profit
	USA	237000000	2009	7.9	760.51	237.00	523.51
	USA	150000000	2015	7	652.18	150.00	502.18
	USA	200000000	1997	7.7	658.67	200.00	458.67
	USA	11000000	1977	8.7	460.94	11.00	449.94
	USA	10500000	1982	7.9	434.95	10.50	424.45
	USA	220000000	2012	8.1	623.28	220.00	403.28
	USA	45000000	1994	8.5	422.78	45.00	377.78
	USA	115000000	1999	6.5	474.54	115.00	359.54
	USA	185000000	2008	9	533.32	185.00	348.32
	USA	78000000	2012	7.3	408.00	78.00	330.00

- Then I sorted the data using the profit column in descending order. And from there I got the top highest profitable movie where Avatar tops the list with 523.51million profit.

movie_title	num_voted_u	num_user_for_re	langua	country	budget	title_y	imdb_sc	gross(in millio	budget(in millior	Profit
Avatar	886204	3054	English	USA	237000000	2009	7.9	760.51	237.00	523.51
Jurassic World	418214	1290	English	USA	150000000	2015	7	652.18	150.00	502.18
Titanic	793059	2528	English	USA	200000000	1997	7.7	658.67	200.00	458.67
Star Wars: Episode IV - A New Hope	911097	1470	English	USA	11000000	1977	8.7	460.94	11.00	449.94
E.T. the Extra-Terrestrial	281842	515	English	USA	10500000	1982	7.9	434.95	10.50	424.45
The Avengers	995415	1722	English	USA	220000000	2012	8.1	623.28	220.00	403.28
The Lion King	644348	656	English	USA	45000000	1994	8.5	422.78	45.00	377.78
Star Wars: Episode I - The Phantom Menace	534658	3597	English	USA	115000000	1999	6.5	474.54	115.00	359.54
The Dark Knight	1676169	4667	English	USA	185000000	2008	9	533.32	185.00	348.32
The Hunger Games	701607	1959	English	USA	78000000	2012	7.3	408.00	78.00	330.00

- For the chart, I selected both profit and budget column and then inserted a scattered chart.



From the above plot the following outliers have been observed

12215.50, 12213.30

4200, -4199.79

2400, -2397.70

1100, -1099.56

- The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***Movies_with_highest_profit.***

Task 3: Find IMDB top 250 movies

In task 3, there are two sections.

For 1st section, I had to find top 250 movies and arrange them according to their ranks.

- At first, I attached a filter on num_voted_users where the values should be greater than 25000.
- Then, I arranged the imdb_score in descending order.
- Further, I selected and copied top 250 movies and pasted them in a new sheet.
- Also, I added the filter option in all the headers.
- For creating rank, I used the following formula =RANK(O2,$O$2:$O$251,0)+COUNTIF($O$2:O2,O2)-1
- Then I used the hide feature to hide the unnecessary columns.
-

movie_title	imdb_sco	Rank
The Shawshank Redemption	9.3	1
The Godfather	9.2	2
The Dark Knight	9	3
The Godfather: Part II	9	4
Pulp Fiction	8.9	5
The Lord of the Rings: The Return of the King	8.9	6
Schindler's List	8.9	7
The Good, the Bad and the Ugly	8.9	8
Inception	8.8	9
Fight Club	8.8	10
Forrest Gump	8.8	11
The Lord of the Rings: The Fellowship of the Ring	8.8	12
Star Wars: Episode V - The Empire Strikes Back	8.8	13
The Matrix	8.7	14
The Lord of the Rings: The Two Towers	8.7	15
Star Wars: Episode IV - A New Hope	8.7	16
Goodfellas	8.7	17
One Flew Over the Cuckoo's Nest	8.7	18
City of God	8.7	19
Seven Samurai	8.7	20

movie_title	imdb_sco	Rank
Se7en	8.6	21
Interstellar	8.6	22
The Silence of the Lambs	8.6	23
Saving Private Ryan	8.6	24
American History X	8.6	25
The Usual Suspects	8.6	26
Spirited Away	8.6	27
Modern Times	8.6	28
The Dark Knight Rises	8.5	29
Gladiator	8.5	30
Django Unchained	8.5	31
The Departed	8.5	32
Memento	8.5	33
The Prestige	8.5	34
The Green Mile	8.5	35
Terminator 2: Judgment Day	8.5	36
Back to the Future	8.5	37
Raiders of the Lost Ark	8.5	38
The Lion King	8.5	39
Alien	8.5	40

movie_title	imdb_sco	Rank
The Pianist	8.5	41
Apocalypse Now	8.5	42
Psycho	8.5	43
Whiplash	8.5	44
The Lives of Others	8.5	45
Samsara	8.5	46
Children of Heaven	8.5	47
American Beauty	8.4	48
Braveheart	8.4	49
WALL·E	8.4	50
Star Wars: Episode VI - Return of the Jedi	8.4	51
Reservoir Dogs	8.4	52
Requiem for a Dream	8.4	53
Amélie	8.4	54
The Other Dream Team	8.4	55
Aliens	8.4	56
Oldboy	8.4	57
Princess Mononoke	8.4	58
Once Upon a Time in America	8.4	59
Lawrence of Arabia	8.4	60

movie_title	imdb_sco	Rank
Das Boot	8.4	61
A Separation	8.4	62
Baahubali: The Beginning	8.4	63
Batman Begins	8.3	64
Inglourious Basterds	8.3	65
Eternal Sunshine of the Spotless Mind	8.3	66
Up	8.3	67
Toy Story	8.3	68
Good Will Hunting	8.3	69
Snatch	8.3	70
Toy Story 3	8.3	71
Scarface	8.3	72
Indiana Jones and the Last Crusade	8.3	73
2001: A Space Odyssey	8.3	74
L.A. Confidential	8.3	75
Monty Python and the Holy Grail	8.3	76
Inside Out	8.3	77
Unforgiven	8.3	78
Amadeus	8.3	79
Downfall	8.3	80

movie_title	imdb_sco	Rank
Hoop Dreams	8.3	81
Raging Bull	8.3	82
The Sting	8.3	83
No End in Sight	8.3	84
Some Like It Hot	8.3	85
The Hunt	8.3	86
Room	8.3	87
Metropolis	8.3	88
V for Vendetta	8.2	89
The Wolf of Wall Street	8.2	90
Finding Nemo	8.2	91
A Beautiful Mind	8.2	92
Die Hard	8.2	93
Gran Torino	8.2	94
The Big Lebowski	8.2	95
How to Train Your Dragon	8.2	96
Trainspotting	8.2	97
Pan's Labyrinth	8.2	98
Blade Runner	8.2	99
Into the Wild	8.2	100

movie_title	imdb_sco	Rank
Lock, Stock and Two Smoking Barrels	8.2	101
Casino	8.2	102
Warrior	8.2	103
Captain America: Civil War	8.2	104
The Thing	8.2	105
Gone with the Wind	8.2	106
The Act of Killing	8.2	107
Howl's Moving Castle	8.2	108
The Bridge on the River Kwai	8.2	109
The Secret in Their Eyes	8.2	110
On the Waterfront	8.2	111
Incendies	8.2	112
The Avengers	8.1	113
Pirates of the Caribbean: The Curse of the Black Pearl	8.1	114
Shutter Island	8.1	115
Kill Bill: Vol. 1	8.1	116
The Sixth Sense	8.1	117
Guardians of the Galaxy	8.1	118
The Truman Show	8.1	119
Sin City	8.1	120

movie_title	imdb_sco	Rank
Jurassic Park	8.1	121
No Country for Old Men	8.1	122
The Terminator	8.1	123
Monsters, Inc.	8.1	124
Donnie Darko	8.1	125
Gone Girl	8.1	126
Mad Max: Fury Road	8.1	127
The Bourne Ultimatum	8.1	128
Million Dollar Baby	8.1	129
Deadpool	8.1	130
The Grand Budapest Hotel	8.1	131
The Martian	8.1	132
The Imitation Game	8.1	133
12 Years a Slave	8.1	134
Groundhog Day	8.1	135
The Revenant	8.1	136
Prisoners	8.1	137
Rocky	8.1	138
There Will Be Blood	8.1	139
The Help	8.1	140
Rush	8.1	141
The Princess Bride	8.1	142
The Wizard of Oz	8.1	143
Platoon	8.1	144
Stand by Me	8.1	145
Hotel Rwanda	8.1	146
Woodstock	8.1	147
Spotlight	8.1	148
Annie Hall	8.1	149
Before Sunrise	8.1	150

movie_title	imdb_sco	Rank
Amores Perros	8.1	151
Butch Cassidy and the Sundance Kid	8.1	152
Nothing But a Man	8.1	153
In the Shadow of the Moon	8.1	154
Akira	8.1	155
Elite Squad	8.1	156
The Celebration	8.1	157
The Sea Inside	8.1	158
The Best Years of Our Lives	8.1	159
Tae Guk Gi: The Brotherhood of War	8.1	160
Slumdog Millionaire	8	161
Black Swan	8	162
District 9	8	163
Catch Me If You Can	8	164
X-Men: Days of Future Past	8	165
Kill Bill: Vol. 2	8	166
Star Trek	8	167
The King's Speech	8	168
The Incredibles	8	169
Ratatouille	8	170
Casino Royale	8	171
Life of Pi	8	172
Jaws	8	173
Blood Diamond	8	174
Shaun of the Dead	8	175
Rain Man	8	176
Her	8	177
The Perks of Being a Wallflower	8	178
Big Fish	8	179
Mystic River	8	180

movie_title	imdb_sco	Rank
The Pursuit of Happyness	8	181
Dallas Buyers Club	8	182
In Bruges	8	183
The Exorcist	8	184
Dead Poets Society	8	185
Boyhood	8	186
Aladdin	8	187
Serenity	8	188
Magnolia	8	189
Mulholland Drive	8	190
The Artist	8	191
Dances with Wolves	8	192
Before Sunset	8	193
True Romance	8	194
Brazil	8	195
Cinderella Man	8	196
The Sound of Music	8	197
A Fistful of Dollars	8	198
The Iron Giant	8	199
Bowling for Columbine	8	200
JFK	8	201
Young Frankenstein	8	202
Dancer in the Dark	8	203
Sling Blade	8	204
Persepolis	8	205
My Name Is Khan	8	206
Sicko	8	207
The Straight Story	8	208
Doctor Zhivago	8	209
Waltz with Bashir	8	210

movie_title	imdb_sco	Rank
Blood In, Blood Out	8	211
Fiddler on the Roof	8	212
Central Station	8	213
Winged Migration	8	214
Little Miss Sunshine	7.9	215
Hot Fuzz	7.9	216
Captain Phillips	7.9	217
Nightcrawler	7.9	218
E.T. the Extra-Terrestrial	7.9	219
Big Hero 6	7.9	220
The Fighter	7.9	221
The Hateful Eight	7.9	222
Moon	7.9	223
The Wrestler	7.9	224
How to Train Your Dragon 2	7.9	225
The Untouchables	7.9	226
Crouching Tiger, Hidden Dragon	7.9	227
Almost Famous	7.9	228
Boogie Nights	7.9	229
Walk the Line	7.9	230

movie_title	imdb_sco	Rank
Halloween	7.9	231
Hero	7.9	232
The Blues Brothers	7.9	233
Ed Wood	7.9	234
The Insider	7.9	235
Letters from Iwo Jima	7.9	236
Straight Outta Compton	7.9	237
Glory	7.9	238
Before Midnight	7.9	239
Once	7.9	240
Amour	7.9	241
My Fair Lady	7.9	242
Do the Right Thing	7.9	243
The Remains of the Day	7.9	244
The Right Stuff	7.9	245
4 Months, 3 Weeks and 2 Days	7.9	246
The World's Fastest Indian	7.9	247
The Chorus	7.9	248
Nine Queens	7.9	249
Veer-Zaara	7.9	250

- From the above table, it can be seen that The Shawshank Redemption is the highest rated movie of all time.
- The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***IMDb_Top_250***

For 2nd section, I had to filter out top 250 movies which are not in English language and store them in a new sheet.

- Now I copied the dataset and pasted on a new sheet and named it as Top_Foreign_Lang_Flim
- Then, after adding the filter option to all the headers, I filtered out English from the language column.
- Then, using the hide feature, I hid all the unnecessary columns.
- From the below table, it can be concluded that movie named The Good, the bad and the ugly, directed by Sergio Leone, is the top rated film in foreign language.

director_name	actor_1_name	movie_title	num_voted_use	language	country	title_year	Profit	imdb_sco	Rank
Sergio Leone	Clint Eastwood	The Good, the Bad and the Ugly	503509	Italian	Italy	1966	4.90	8.9	8
Fernando Meirelles	Alice Braga	City of God	533200	Portuguese	Brazil	2002	4.26	8.7	19
Akira Kurosawa	Takashi Shimura	Seven Samurai	229012	Japanese	Japan	1954	-1.73	8.7	20
Hayao Miyazaki	Bunta Sugawara	Spirited Away	417971	Japanese	Japan	2001	-8.95	8.6	27
Florian Henckel von Donnersmarck	Sebastian Koch	The Lives of Others	259379	German	Germany	2006	9.28	8.5	45
Ron Fricke	Collin Alfredo St. Dic	Samsara	22457	None	USA	2011	-1.40	8.5	46
Majid Majidi	Bahare Seddiqi	Children of Heaven	27882	Persian	Iran	1997	0.75	8.5	47
Jean-Pierre Jeunet	Mathieu Kassovitz	Amélie	534262	French	France	2001	-43.80	8.4	54
Chan-wook Park	Min-sik Choi	Oldboy	356181	Korean	South Korea	2003	-0.82	8.4	57
Hayao Miyazaki	Minnie Driver	Princess Mononoke	221552	Japanese	Japan	1997	-2397.70	8.4	58
Wolfgang Petersen	Jürgen Prochnow	Das Boot	168203	German	West Germany	1981	-2.57	8.4	61
Asghar Farhadi	Shahab Hosseini	A Separation	151812	Persian	Iran	2011	6.60	8.4	62
S.S. Rajamouli	Tamannaah Bhatia	Baahubali: The Beginning	62756	Telugu	India	2015	-11.53	8.4	63
Oliver Hirschbiegel	Thomas Kretschmann	Downfall	248354	German	Germany	2004	-8.00	8.3	80
Thomas Vinterberg	Thomas Bo Larsen	The Hunt	170155	Danish	Denmark	2012	-3.19	8.3	86
Fritz Lang	Brigitte Helm	Metropolis	111841	German	Germany	1927	-5.97	8.3	88
Guillermo del Toro	Ivana Baquero	Pan's Labyrinth	467234	Spanish	Spain	2006	24.12	8.2	98
Joshua Oppenheimer	Anwar Congo	The Act of Killing	23836	Indonesian	UK	2012	-0.52	8.2	107
Hayao Miyazaki	Christian Bale	Howl's Moving Castle	214091	Japanese	Japan	2004	-19.29	8.2	108
Juan José Campanella	Ricardo Darín	The Secret in Their Eyes	131831	Spanish	Argentina	2009	18.17	8.2	110
Denis Villeneuve	Lubna Azabal	Incendies	80429	French	Canada	2010	0.06	8.2	112
Alejandro G. Iñárritu	Adriana Barraza	Amores Perros	173551	Spanish	Mexico	2000	3.38	8.1	151
Katsuhiro Ôtomo	Mitsuo Iwata	Akira	106160	Japanese	Japan	1988	-1099.56	8.1	155
José Padilha	Wagner Moura	Elite Squad	81644	Portuguese	Brazil	2007	-3.99	8.1	156
Thomas Vinterberg	Ulrich Thomsen	The Celebration	65951	Danish	Denmark	1998	0.35	8.1	157
Alejandro Amenábar	Belén Rueda	The Sea Inside	64556	Spanish	Spain	2004	-7.91	8.1	158
Je-kyu Kang	Min-sik Choi	Tae Guk Gi: The Brotherhood of War	31943	Korean	South Korea	2004	-11.69	8.1	160
Sergio Leone	Clint Eastwood	A Fistful of Dollars	147566	Italian	Italy	1964	3.30	8	198
Vincent Paronnaud	Catherine Deneuve	Persepolis	70194	French	France	2007	-2.86	8	205
Karan Johar	Shah Rukh Khan	My Name Is Khan	69759	Hindi	India	2010	-7.98	8	206
Ari Folman	Ari Folman	Waltz with Bashir	46107	Hebrew	Israel	2008	0.78	8	210
Walter Salles	Fernanda Montenegro	Central Station	28951	Portuguese	Brazil	1998	2.70	8	213
Ang Lee	Chen Chang	Crouching Tiger, Hidden Dragon	217740	Mandarin	Taiwan	2000	113.07	7.9	227
Yimou Zhang	Jet Li	Hero	149414	Mandarin	China	2002	-30.92	7.9	232
Clint Eastwood	Yuki Matsuzaki	Letters from Iwo Jima	132149	Japanese	USA	2006	-5.25	7.9	236
Michael Haneke	Isabelle Huppert	Amour	70382	French	France	2012	-8.67	7.9	241
Cristian Mungiu	Anamaria Marinca	4 Months, 3 Weeks and 2 Days	44763	Romanian	Romania	2007	0.60	7.9	246
Christophe Barratier	Jean-Baptiste Maunier	The Chorus	44151	French	France	2004	-1.87	7.9	248
Fabián Bielinsky	Ricardo Darín	Nine Queens	38215	Spanish	Argentina	2000	-0.28	7.9	249
Yash Chopra	Shah Rukh Khan	Veer-Zaara	34449	Hindi	India	2004	-4.08	7.9	250

- The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***Top_Foreign_Lang_Flim***

Task 4: Find the best directors

- At first, I selected the entire table and created a pivot table.
- Then, I added director_name column to Rows and imdb_score column to Values and changed the imdb_score field from Sum to Average.
- Next, I change the name and sorted Director Name column to top 10.
- Then, I sorted the Average of imdb_score column in descending order and Director Name in ascending order.
- The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***top10directors***

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
Asghar Farhadi	8.4
Marius A. Markevicius	8.4

From the table, it can be concluded that the highest rated director is Charles Chaplin with 8.6 rating.

Task 5: Find popular genres

Here I have used the similar method from the previous task.

- First, I selected the table and created a pivot table.
- Then, I added genres column to rows and values and changes the value field from Sum to Count.
- Then, I sorted the Count of genres in descending order and Genre column in top 10.

- The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***popular genres***

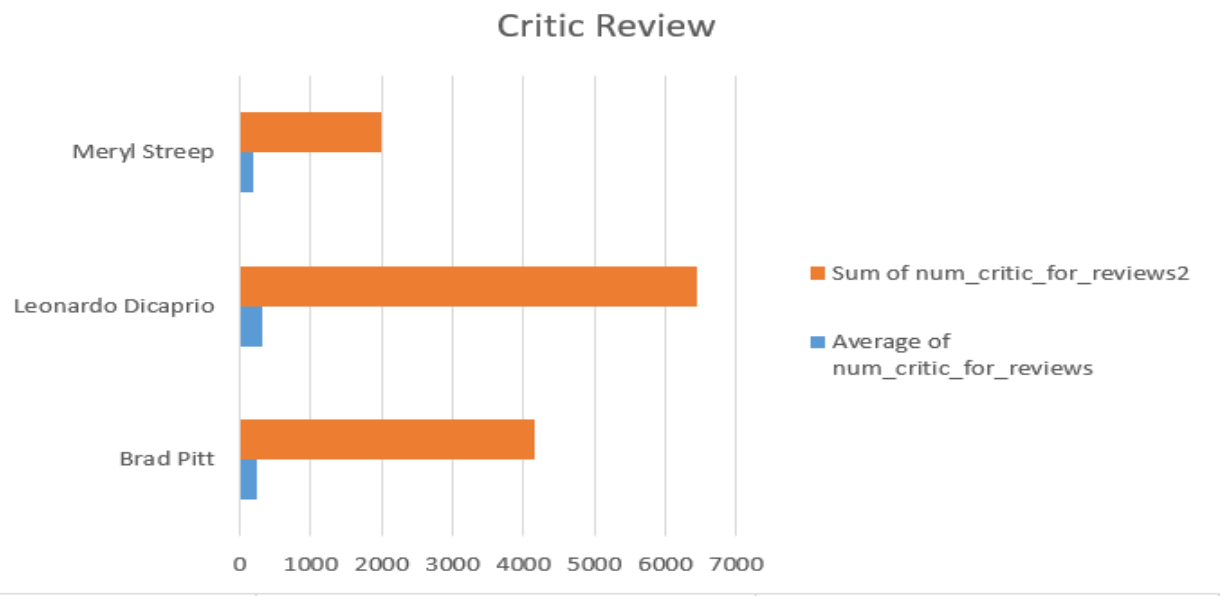
Genre	Count of genres	Average of imdb_score
Drama	152	7.040131579
Comedy|Drama|Romance	149	6.495302013
Comedy|Drama	147	6.583673469
Comedy	145	5.840689655
Comedy|Romance	135	5.896296296
Drama|Romance	118	6.95
Crime|Drama|Thriller	80	6.865
Action|Crime|Thriller	54	6.403703704
Action|Crime|Drama|Thriller	48	6.522916667
Action|Adventure|Sci-Fi	45	6.668888889
Comedy|Crime	45	6.037777778
Grand Total	**1118**	**6.483542039**

From the table it can be concluded that the popular genre is Drama with 152 counts and 7.04 average of imdb score.

Task 6: Find the critic-favourite and audience-favourite actors

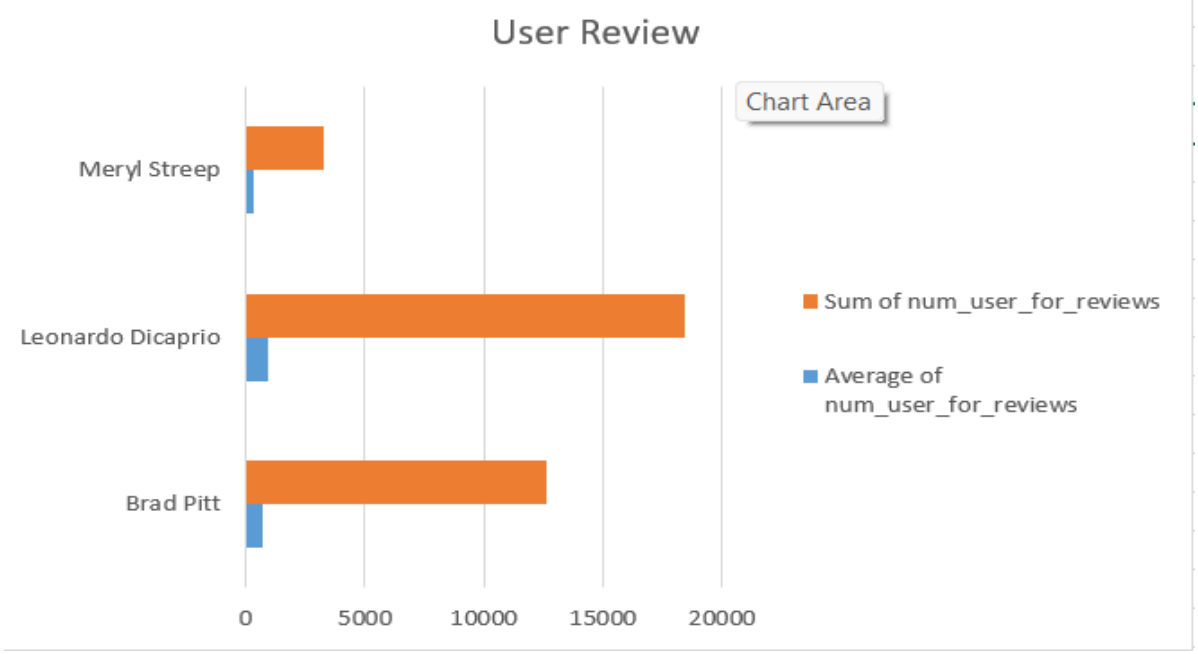
- Firstly, I created three new dataframes for all three separate actors i.e., Meryl Streep, Brad Pitt and Leonardo Dicaprio.
- Then, I created a pivot table and added actor_1_name to the filters section and the rest of the necessary columns to the Rows section.
- Changed the layout into Tabular form
- Now filtered Meryl Steep, Leonardo Di Caprio and Brad Pitt data, copied and pasted their data to the destined sheet.
- Then, I selected the table of Meryl Streep and formed a connection using power query editor. I did the same with others as well.
- Next, to append the data, I moved to Get Data section and combined or appended the data of those three data sheets using power query editor.

Actors	Average of num_critic_for_reviews	Sum of num_critic_for_reviews2
Brad Pitt	245	4165
Leonardo Dicaprio	322.2	6444
Meryl Streep	181.4545455	1996



- Now using the pivot table analysis, I got the following results

Actors	Average of num_user_for_reviews	Sum of num_user_for_reviews
Brad Pitt	742.3529412	12620
Leonardo Dicaprio	922.55	18451
Meryl Streep	297.1818182	3269



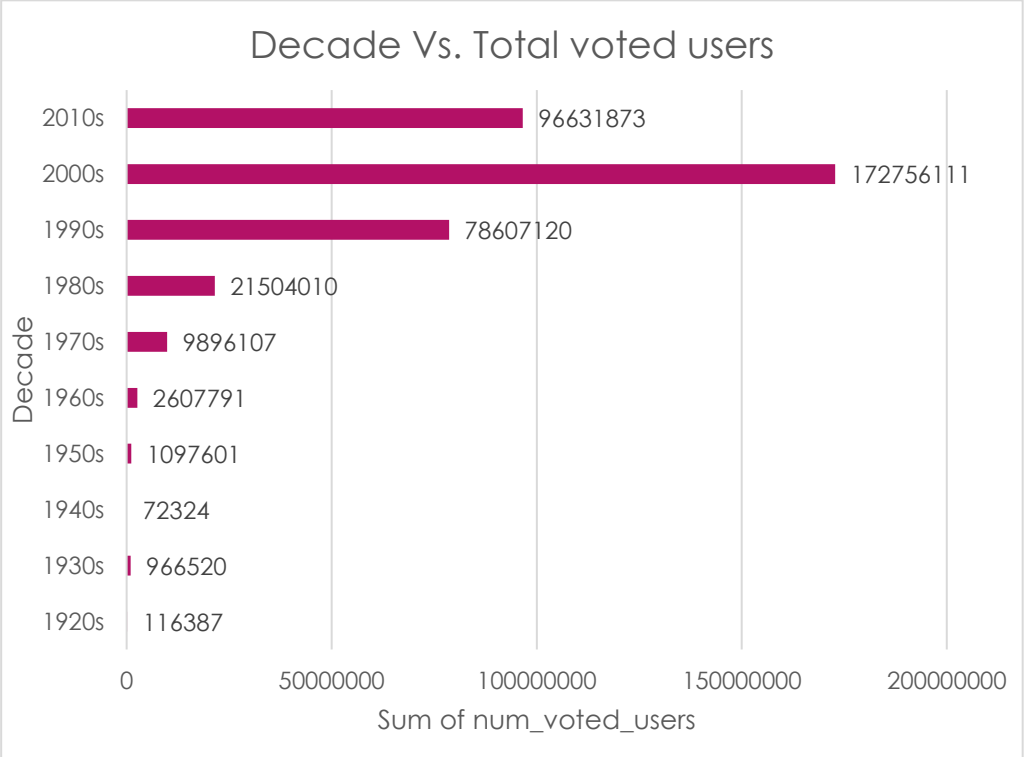
From the above two tables and data, it can be concluded that Leonardo Dicaprio has been critic-favourite as well as audience-favourite.

The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***critic review*** & ***user review***

For the decade dataframe, I did the following steps

- First, I created a copy of the dataset and pasted in a new sheet and named it as decade.
- Then, I opted for a pivot table where I grouped the title_year column and renamed it accordingly.
- Later, I opted for a bar chart.

Decade	Sum of num_voted_users
1920s	116387
1930s	966520
1940s	72324
1950s	1097601
1960s	2607791
1970s	9896107
1980s	21504010
1990s	78607120
2000s	172756111
2010s	96631873
Grand Total	**384255844**



From the above table and bar plot, it can be concluded that the highest voted users were belong to 2000s.

The data can be seen on a new sheet of ***IMDB Data(cleaned).xlsx*** under the name ***df_by_decade***

5. TECH STACK USED

- Microsoft Excel 2019
- Microsoft Word 2019
- Google drive

6. INSIGHTS

After the completion of this project, I understood some advance uses of Microsoft Excel. This project also helped me getting into the world of power query. By accomplishing this project successfully, I understood the need of power query in managing and manipulating the data in Excel. The level of data cleaning and data processing was a bit advanced from the previous project yet very challenging and exciting.

As suggested, I used the Root Cause Analysis developed by Sakichi Toyoda. I have listed the important questions and answers that I have jotted down while attempting this project

- Why the most rated movie is not the most profitable movie?

The IMDB rating is calculated only when someone voted on their portal whereas the profit is calculated on the basis of number of tickets sold. Therefore, the possibility of those people who might have watched the movie but not rated them in IMDB comes to light, hence generating the difference in rating and profits.

- Why most of the movies are in English language?

The native language of USA is English and they are currently the highest earning economy in the world. Therefore, directors and producers in Hollywood have the capability to invest more on a film. And once someone invests that much amount of money, they automatically hires the best artists in the industry. Also, English is the most common language used as communication tool in most of the countries, hence making a movie in this language automatically reaches most of the people worldwide.

- Why drama genre is the dominating ones over others?

The taste of movies varies from person to person as well as well as generation. Therefore, some genres were capable of targeting a specific set of audience. But drama is the only genre that is being enjoyed by anyone. It can be watched by an individual or with

the family as well. The drama genre narrow downs the generational gap hence making it a dominating one.

- Why 2000s received the greatest number of votes?

The year within 2000s saw a great enhancement in the technology sector as well as economical sector. This happened in almost all the countries. Also, the invention of Blu-ray picture format within 2000s shooted up the movie industry. Many countries took this opportunity to remove or shorten the taxes on movies. Thus, making people to go, watch and vote for the movies even more.

7. RESULTS & CONCLUSION

At the end, analyzing movies using data analysis techniques can prove to be incredibly useful for movie makers, investors and stakeholders. While it may not be a concern for the general public, such analysis plays a crucial role during the pre-production and post-production phases of movies. It can help in making informed decisions regarding the casting, budget, genre, and other aspects of movie production.

It is important to note that the highest IMDb rating does not necessarily translate to the highest profit. Profit is calculated based on the number of tickets sold by theaters all over the world. Therefore, it is important to consider both critical and commercial success while analyzing movies.

Furthermore, it is important to understand the audience's preferences when it comes to movie genres. Most people prefer movies with Comedy/Drama genre or a combination of both. On the other hand, movies with Action/Horror genre may not be as popular among audiences. Therefore, directors and production teams must keep these points in mind and perform pre-production analysis before the commencement of filming.

In conclusion, data analysis can provide valuable insights into the movie industry and help in making informed decisions that can ultimately lead to success. By using techniques such as data cleaning, grouping, and visualization, it is possible to extract meaningful information from movie datasets that can guide decision-making in the movie industry.

8. DRIVE LINK

https://drive.google.com/drive/folders/1bwEYoH4snMBDomXyJNwHgbJCxGhSCKd0

Folder Name: IMDB Movie Analysis

Name of the report file: project_5(IMDB movie analysis)

Data File Name: IMDB Data(cleaned).xlsx